PRESIDENCY UNIVERSITY

DEPARTMENT OF LIFE SCIENCES

PhD Thesis Synopsis

Studies on interactions of HIV-1 accessory proteins from different subtypes and their interacting partners

Name of Candidate: **JOYEETA DATTA**

Registration Number with date: **R-18RS18210188 (11th February, 2020)**

Name of Supervisor: **Dr. KALYAN GIRI**

Name: JOYEETA DATTA

Supervisor: Dr. KALYAN GIRI

Co-supervisor (if any):

Title of Thesis: Studies on interactions of HIV-1 accessory proteins from different subtypes and their interacting partners

PhD registration Number with date: R-18RS18210188 (11th February, 2020)

Enrolment Number with date: PU/PHD/DLS/2018/105 (30th October, 2018)

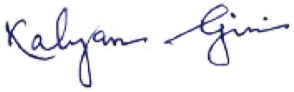
Place of work: Presidency University, Kolkata

Residential Address: 77/26, Rammohan Place, Konnagar, Hooghly-712235

Corresponding Address: Do

Email: mmdtt26@gmail.com, joyeeta1.rs@presiuniv.ac.in

Phone Number: 7003597234/8981258393



Signature of Research Scholar **Signature of Supervisor**

CONTENTS	PAGE NUMBER
ABSTRACT	4
INTRODUCTION	5-6
OBJECTIVES	6
METHODOLOGY	7-15
RESULTS	16-55
OBJECTIVE 1	16-33
OBJECTIVE 2	34-42
OBJECTIVE 3	43-55
CONCLUSION	56-57
SIGNIFICANCE OF THE STUDY	58
FUTURE DIRECTION	59
REFERENCE	60-68

ABSTRACT

HIV-1, the major causative agent of AIDS results in significant number of deaths globally. Novel drug development for HIV-1 is challenging due to the presence of multiple subtypes. In this study, the similarity and dissimilarity between the subtypes have been studied according to the accessory proteins (Nef, Vif, Vpr and Vpu) of the virus. Several reports suggest the importance of these accessory proteins in viral disease progression and pathogenesis. Initial finding of sequence analysis data elucidates the sequence variation between the subtypes of HIV-1. In spite of having similar secondary structural motifs, distinct differences have been observed in the tertiary structures across the subtypes. Changes in dynamics of druggable pocket of Nef protein has been determined while Vif protein has the uniform druggable pocket region. Vpr and Vpu proteins show differences in probe binding cluster region having identical pocket regions across the subtypes. Additionally, different molecules were generated against the same pocket for different subtypes of Vpr protein. However, these molecules are seen to have almost similar molecular contact with the protein from different subtypes. On the other hand, subtypic differences have been spotted out in case of Nef protein having variations in binding poses of a particular molecule. Notably, these accessory proteins are seen to maintain the similar interaction site and molecular contact with the host protein partner across the subtypes. In conclusion it can be said that the accessory proteins behave similarly in case of large molecular interactions like protein-protein interaction but may differ in case of small molecule like drug or inhibitor binding. This study discusses the similarity and dissimilarity present across the subtypes of HIV-1 from sequence to structure levels. It emphasizes on the druggable pocket, the small molecule binding pattern as well as protein partner interaction pattern. All these findings may help to get the idea whether HIV-1 treatment needs a subtype specific drug approach or any general inhibitor would be applicable for all the subtypes.

INTRODUCTION

Human Immunodeficiency Virus (HIV) has been a continuum global health challenges since its discovery. It attacks the immune system specially the CD4, MHC cells and diminishes the cell counts in the body which in progression makes a person susceptible to infections [1]. Gradually, HIV infection progress to Acquired Immunodeficiency Syndrome (AIDS) with severely damaged immune system. The virus transmitted by the direct contact of body fluid like semen, vaginal fluid, blood, saliva, rectal fluid of any HIV infected person to the others. HIV can be transmitted through the amniotic fluid and breast milk from the infected mother to the child [2].

Between two major types of HIV, HIV-1 is the most prevalent being responsible for majority of infection worldwide whereas, HIV-2 occurs in fewer number of individuals especially in West Africa. The complicated transmission efficiency and less virulence of HIV-2 restricted its geographic distribution compared to HIV-1. A closely related species Simion Immunodeficiency Virus (SIV) that affects the primates, chimpanzees, monkeys and apes. HIV-1 can be classified in four phylogenetic groups: Group Major (M), Group O (outlier: only seen in west central Africa), Group N (Non-M: only seen in Cameroon) and Group P. Group M is further classified in to different subtypes, sub-subtypes, circulatory recombinant forms (CRFs) and unique recombinant forms (URFs). Group M causes the majority of HIV-1 infections worldwide and can be broadly classified in to 9 different subtypes: A, B, C, D, F, G, H, J and K based on the genetic difference. Article suggested that genetic differences across the subtypes is about 25-35% whereas within the subtypes is 15-20% [3-7]. There are differences in viral disease progression, pathogenesis as well as in drug resistance between the subtypes [8-12].

HIV-1 is a member of group of Lentivirus within the family of Retroviridae and subfamily Orthoretroviridae. The genome consists of two identical positive sense single-stranded RNA molecules enclosed within a capsid which gets surrounded by a plasma membrane originated from host cell containing the viral envelope proteins [13]. The spike glycoprotein formed by the envelope protein required for infection to the host cell. With a long terminal repeat (LTR), the genome codes for 9 proteins of which three are structural proteins (Gag, Pol and Env), Two regulatory proteins (Tat and Rev) and four accessory proteins (Vif, Vpr, Vpu and Nef) [14,15]. The accessory proteins don't take part in active viral replication procedure but several studies demonstrate that in the absence of aborted or full-length accessory proteins, the virus loses its infection capability. They play crucial roles in viral replication cycle and elucidate several immune responses [16,17]. Viral Infectivity Factor (Vif) is a 192 aa long protein which

encounters and degrades different APOBEC3 that induces hyper-mutation in viral DNA [18,19]. Viral Protein R (Vpr), a 96aa long multifunctional protein which facilitates the nuclear import of the viral pre-integration complex (PIC), G2 cell cycle arrest, apoptosis and cytotoxicity. It interacts with numerous host proteins such as DCAF1 (DDB1-CUL4-associated factor-1), an E3 ubiquitin ligase which degrades different proteins and promotes G2-cell cycle arrest [20-25]. Viral Protein U (Vpu), the smallest 82 aa long accessory proteins which induces the virion release from infected cells by interaction with BST2 (tetherin) and downregulation of CD4 cells [26]. Negative Factor (Nef), the largest accessory protein (208aa) modulates various cellular processes to enhance viral replication and evade immune responses. It interacts with several members of the SRC family kinases (HCK, Lyn ad Lck) and promotes downregulation of various immune cells like MHC-I, CD4 [27].

HIV-1 causes almost 85% of HIV infection worldwide however there is inadequacy of therapeutics for HIV-1 treatment. There is always a constant requirement of new drug as the conventional drugs become functionally ineffective over the time. High mutation rate promotes genetic diversity in HIV-1 which gives rise to multiple subtypes or clades. Moreover, majority of available drugs (NRTIs, NNRTIs, PIs and INSTIs) are designed against the structural proteins and considering the subtype B specially [28]. Several reports suggest that drug resistance has been shown to non-B subtypes by the effect of subtype specific polymorphism and cross resistance [29,30].

The accessory proteins are multifunctional and possess various host-virus protein-protein interactions. Recent days the accessory proteins become interesting field for researchers though very few information is available. Hence, we want to design our study emphasized on the subtype specific differences starting from sequence level to structural level of the accessory proteins.

OBJECTIVES

1. Investigation of sequence variation and structural alteration of HIV-1 accessory proteins from different subtypes.
2. Identification of different ligand binding pockets and druggability of HIV-1 accessory proteins.
3. Characterization of protein-protein interaction of HIV-1 accessory proteins with the host binding partners and identifying interaction hot-spot regions.

METHODOLOGY

1. Protein sequence retrieval

Sequence of four accessory proteins were retrieved from the Los Alamos database [31]. The sequences were segregated in to subtypes, country or even different year of reporting. Side by side the reviewed sequences of four accessory proteins were downloaded from UniProt Database and segregated in to different subtypes [32].

The sequence strength in Los Alamos database is almost 10 folds higher than the UniProt database. However, the sequences contain unrecognised amino acids in many cases where as UniProt has reviewed sequences which is more reliable to use. Hence, for the sequence related research Los Alamos sequences has been used but for structure related study UniProt sequences are used. Pairwise sequence alignment of respective set of protein sequences from Los Alamos and UniProt protein database show 99% similarity in each case. Total ten subtypes (A, B, C, D, F1, F2, G, H, J &K) are being for Vif, Vpr and Nef. Only 6 subtypes were reported for Vpu proteins except (F2, G, J and K not reported in UniProt).

2. Consensus sequence determination

Multiple sequence alignment (MSA) was executed in multiple stages to determine the consensus sequences of each subtype. There is a chance of losing data in case to executing MSA with this large number of data hence multiple stages of MSA needs to be executed (**Figure 1**). Jalview software is used to determine the consensus sequences [33,34].

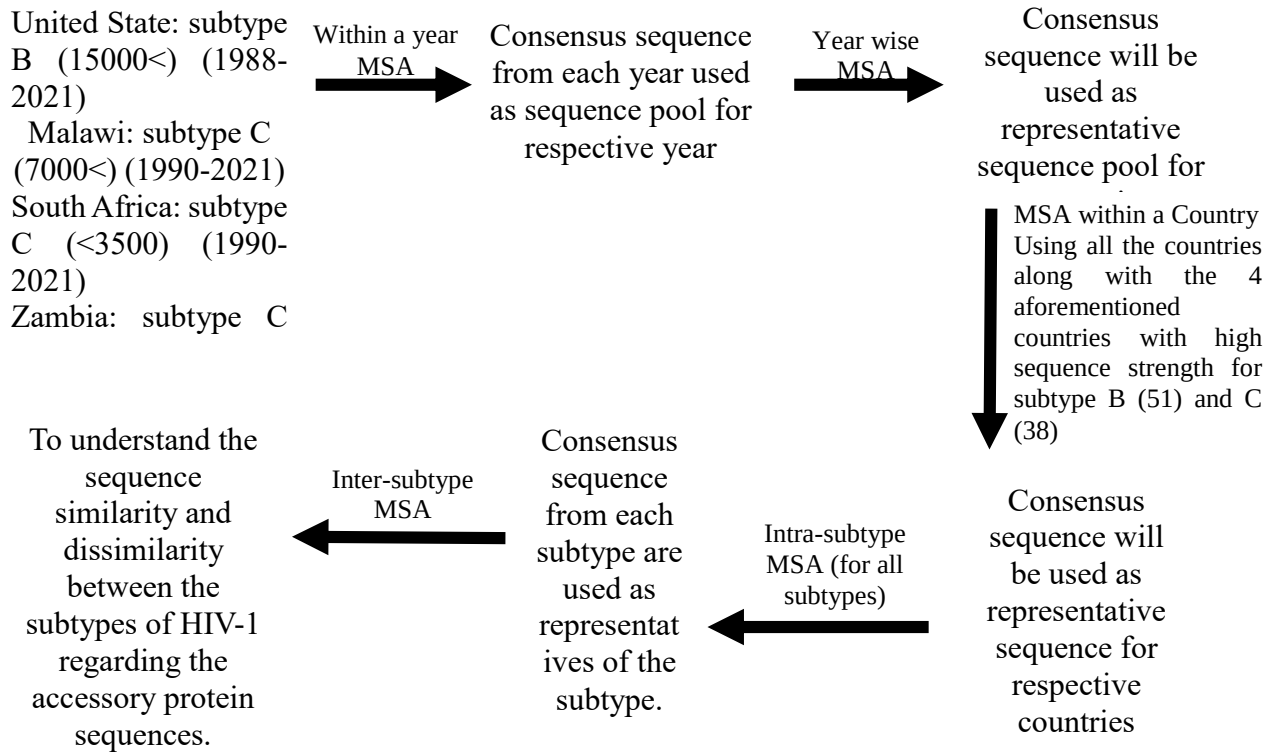

United State: subtype B (15000<) (1988-2021)
Malawi: subtype C (7000<) (1990-2021)
South Africa: subtype C (<3500) (1990-2021)
Zambia: subtype C
Within a year MSA
Consensus sequence from each year used as sequence pool for respective year
Year wise MSA
Consensus sequence will be used as representative sequence pool for
MSA within a Country Using all the countries along with the 4 aforementioned countries with high sequence strength for subtype B (51) and C (38)
Consensus sequence will be used as representative sequence for respective countries
Intra-subtype MSA (for all subtypes)
Consensus sequence from each subtype are used as representatives of the subtype.
Inter-subtype MSA
To understand the sequence similarity and dissimilarity between the subtypes of HIV-1 regarding the accessory protein sequences.

Figure 1: The process for determine the consensus sequences from large sequence pool of each accessory proteins reported in Los Alamos database. MSA within a year reports the consensus sequence of the respective year for US:B, ME:C, ZA:C and ZM:C. MSA with the sequences from all the year provides the sequence pool for the mentioned countries for subtype B and C. MSA with the sequences with in a country determine the consensus for each country. Intra-subtype MSA with the consensus from each country provides the idea sequence variability in one subtype. Consensus sequence from the intra-subtypes data from all the ten subtypes determines the conservation and variability across the subtypes of a particular accessory protein.

3. Clustering of protein sequence

Clustering of protein sequence helps to identify sequences with high degree of similarity [35]. Here, the sequences of accessory proteins were clustered with 75% similarity cut off for both across and between the subtypes. A low similarity cut off will help to categorise the sequences in group with high stringency. Inhouse Biopython scripts was used for the clustering.

4. Shannon entropy calculation

Shannon entropy is a measurement of the randomness of a set of data. For protein sequence data set, it quantifies the diversity of amino acids at each position [36]. The entropy (H) for a position in the sequence can be calculated using the following formula:

$$H = -\sum_i pi \log_2(pi)$$

$$pi = \text{the probability of the ith amino acid at that position}$$

Shannon entropy (Los Alamos site) of both intra-subtype and inter-subtype protein sequences have been calculated.

5. Determination of secondary structure

Secondary structure of a protein is the local folded structure forms by backbone interaction of polypeptide. The secondary structures for 10 subtypes of four accessory proteins of HIV-1 was determined individually using PSIPRED Tool [37]. It helps to find the conservancy in the secondary structures of the protein across the subtypes.

6. Model structure preparation

Model structures of accessory protein of ten HIV-1 subtypes were prepared using HOMOLOGY MODELLING (SWISS-MODEL) [38]. After selecting the appropriate PDB as the template file, the model with the highest GMQE score has been taken as the representative model structure of any particular subtype of any particular protein.

Ramachandran Plot was computed for the backbone torsional angle of the protein structures to validate the model structures to validate the models [39].

7. Prediction of Druggable pocket

Druggable pocket is the specific site of a protein which has potential to bind small molecules. It is one of the important aspects in the field of drug discovery study. Here, the similarities and variations of druggable pockets across the subtypes for a particular accessory protein were studied [40,41]. Total 5 different web servers were used to predict the druggable pockets present in a particular protein for all the 10 subtypes.

1. Fpocket [42,43]
2. FpocketWeb [44]
3. DogSiteScorer [45,46]
4. DeepSite [47]
5. CB-Dock cavity detection [48-50]
6. FtMap [51-54]

8. *De novo* Drug Designing

AutoGrow4 follows genetic algorithm to generate novel small molecules by fragment-based drug design (FBDD) method [55]. The initial population starts with randomly combined fragments which creates the library of seed molecules. These seed molecules are then docked into a user-specified protein pocket, scored, and ranked by their calculated fitness using the AutoDock VINA and the poses are ranked based on their calculated fitness [56]. The top-scoring molecules (based on the VINA score expressed in kcal/mol) from the previous generation are then used to seed new generations. Then, it follows an evolutionary cycle: selection of top scored molecules as the seed for the next generation (selection), or Two seed molecules are combined to provide new molecules (crossover) or random modification of seed molecules to generate new molecules (mutation). This evaluation, scoring and evolutionary cycle repeat for multiple generations to generate optimized novel molecules. The ZINC15 database is used as the default fragment library, which is divided into four groups according to molecular weight (MW $\leq$ 100 Da, 100 Da < MW $\leq$ 150 Da, 150 Da < MW $\leq$ 200 Da, and 200 Da < MW $\leq$ 250 Da) [57]. OpenBabel 2.4.1 is utilized to convert the SMILES fragment libraries into 3D pdbqt format, adding hydrogen atoms and optimizing the intramolecular hydrogen-bond network (at pH 7.4) [58]. RobustRxn set, a library of 58 reactions has been set for use and generation of molecules. A specific pocket (box size 44Å*44Å*44Å) of HIV-1 Vpr protein of 10 subtypes is provided with the number of generations set to 30 for MW of 100-150 Da and docking exhaustiveness set to 15. The top scored molecules are then evaluated for further validation and studies.

9. Molecular Docking studies

Protein-protein and protein-ligand dockings are executed to investigate the stability, binding pose and molecular contact in the complexes between the receptor and the donor **(Table 1&2)**. Based on different scoring algorithm the servers or software determine some score depending upon which the best docked complex is being chosen.

Table1: Protein-ligand docking [59]: In this docking the interaction between receptor protein and small molecule ligand is studied. The binding poses and binding affinity between the protein and ligand are studied to find out the binding similarity and dissimilarity across the subtype.

Protein	Ligand	Web Server/Software
Vpr proteins from 10 subtypes	1. Molecules showing *in-vitro* anti-Vpr activity from various reports [60-62]	AutoDock Vina [63,64]: It is a docking tool that evaluates the binding pose and affinity of ligands in a predictive protein binding site. Parameter Grid box size = (44*44*44) A^3 Exhaustiveness = 32 AutoDock 4.0 [65] : to confirm the exact binding pose
	2. Damnacanthal and Hematoxyline derivative (Based on having highest binding affinity with all the subtypes)	AutoDock Vina [63,64] Parameter Grid box size = (44*44*44) A^3 Exhaustiveness = 32
	3. AutoGrow generated Molecules (L9 and L10, Based on Vina energy and ADMET properties)	
Cluster candidates from Nef protein of 10 HIV-1 subtypes	1. FDA approved 1400 molecules from ZINC20 database [66]	AutoDock Vina Parameter: (24*24*40) A^3 Exhaustiveness = 32
	2. 4 selected molecules from previous step (based on Vina energy)	CB-DOCK2 [50]

Table 2: Protein-protein docking: In this docking the interaction between receptor proteins and their host protein partner is studied to check the binding poses between the host and Virus proteins.

Proteins	Host protein	Server
Nef proteins from 10 HIV-1 subtypes	SH3 complex(4u5w)	Z-Dock [67,68]
Vif proteins from 10 HIV-1 subtypes	Apobec3G, Apobec3F and APobec3H	Cluspro [69] and Z-Dock
Vpu proteins from 6 HIV-1 subtypes (excluding F2, G, J and K)	BST2(Tethrine)	HADDOCK 2.4 [70,71]

10. Molecular dynamics (MD) simulation

This method is used for different purposes in this study. The table contains the full list of simulations done with applied force fields, water model, size and shape of simulation box, simulation time (**Table 3**).

Nef and Vpr protein: every system was neutralized with an appropriate number of Na+ and Cl- ions, followed by 5000 steps of energy minimization using the Steepest Descent algorithm [72]. Subsequently, the energy-minimized systems were subjected to NVT and NPT equilibration steps (500 ps each), with a positional constraint of 1 kcal/mol/$Å^2$ applied to protein heavy atoms at 300 K and 1 bar pressure. The NVT ensemble utilized the V-rescale (modified Berendsen temperature) coupling method, while the NPT ensemble employed the Parrinello-Rahman barostat method [73, 74]. Particle Mesh Ewald (PME) summation method was employed for long-range electrostatic interactions [75]. Nef protein has been seen to have higher sequence variation in both intra and inter subtype. Hence, three independent MD simulation has been run which increase the stringency of the protein stability investigation. This is also required for the identification of distinct conformations of Nef protein. Whereas, Vpr protein has less sequence variation across the subtypes, hence, one set of simulation has been executed in two different force fields (Amber and Charmm) to investigate the protein dynamicity. Similarly, one simulation remained enough for the investigation protein-ligand complex stability (result section).

Vpu protein: Transmembrane protein-protein complexes are prepared by Charmm-GUI webservers [76].

Table 3: Molecular dynamics simulation performed in this study. Gromacs 2020.5 suite is used for each simulation run [77,78]. Ligand parameters were generated using Acpype 3.0 [79].

Accessory Protein		Purpose of the MD simulation	Force Field/Water Model [80]	Simulation box shape/size	Time (ns)	Repetition
Nef	10 subtypes	Dynamic stability checking	Charmm22 [81] /TIP3P	Cubic	50	Triplicate each
Vpr	10 subtypes	Dynamic stability checking of protein	Charmm22 [81] /TIP3P	Dodecahedron/10 Å	50	Single
		Dynamic stability checking of pocket region	Amber99sb [82]-ILDN/TIP3P	Dodecahedron/10 Å	50	
	10 subtypes With ligand 1.Damnacanthal 2. Hematoxyline derivative 3. AutoGrow generated novel inhibitor-like molecules (L9 and L10)	Dynamic stability of Protein-Ligand complexes	Charmm22/TIP3P	Dodecahedron/8 Å	100	Single
			AmberFf99sb-ILDN/TIP3P	Dodecahedron/10 Å	100	
Vpu	10 subtypes (TM domain) with protein BST2	Dynamic stability of protein-protein complexes	System prepared by Charmm-Gui software Water model TIP3P	DOPC Membrane [83]	10	Single

11. Post Simulation analysis

Calculation of Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Radius of gyration (R_g), Principal Component Analysis (PCA), Binding free energy calculations (MM/PBSA and MM/GBSA) are done with the trajectory of the MD simulations. The details are given in **Table 4**.

Table 4: Post Simulation analysis of the aforementioned simulated systems. These analysis determine the dynamic stability, different conformation from clusters, the ligand stability in APO system.

Protein	Post Simulation Analysis
Nef	RMSD: To evaluate Dynamic change of Protein in simulation trajectory. (Protein backbone specific) Using gmx_cluster command and gromos clustering method Cluster analysis (cut off 0.25Å): To identify different possible conformations by clustering the whole trajectory (triplicate applied).
Vpr	RMSD: To evaluate Dynamic change of full and truncated Protein (helix region) in simulation trajectory. (Protein backbone specific) RMSF: To evaluate fluctuation of the pocket (helix region)(Cα specific)
	PCA [84,85]: To determine most distinctive motion of protein backbone by setting MD trajectory into smaller dimension. (first two principal components selected) Free energy surface calculation and energies were weighted. Post RMSD calculation of lowest energy conformations were determined.

	Vpr-Damnacanthal And Vpr-hematoxyline derivative	Using Gromcas Command RMSD: To evaluate the protein-ligand complexes in simulation trajectory. (ligand heavy atom specific) RMSF: To compare the fluctuation of protein in Apo and Holo state. R_g: To quantify the protein compactness. Using g_mmpbsa [86,88] Binding free energy calculation: Molecular mechanics Poisson-Boltzman Surface Area (MMPBSA) was calculated for the protein-ligand binding free energy. (last 10ns considered)
	Vpr- L9 and L10	Using Gromcas Command RMSD: To evaluate the protein-ligand complexes in simulation trajectory. Using gmx_MMPBSA command Binding free energy calculation: Molecular mechanics Generalised Born Surface Area (MMGBSA) was calculated for the protein-ligand binding free energy [89-91]. Binding free energy decomposition: to evaluate the total energy profile of individual amino acids (distance between protein-ligand was set to 4Å) Using gmx_hbond command Hydrogen bonding: To quantify the number of hydrogen bond between protein ligand in last 10ns.
Vpu	Vpu-BST2	RMSD: to evaluate the dynamic stability of protein-protein complexes.

RESULTS

OBJECTIVE 1

The purpose of this objective is to search for the similarity and dissimilarity across the subtypes of HIV-1 on the basis of the accessory proteins data (**Table 5 &6**). Subtype B and C are seen to have highest abundancy followed by subtype A and D. Rest of the subtypes are generally seen to be clustered in specified geographical areas: example subtype F2 and K are mainly found in Cameroon and Congo respectively. From the data it can be concluded that African continent has the most diversity while US has reported the largest number of infected population.

Table 5: Sequence strength of four accessory proteins for different subtypes. Subtype A, B and C are the most prevalent types and they are shown to be clustered in some particular area. Subtypes J&K are shown to have least number of sequences. The data of subtype F is stored by F1 and F2 in LANL database.

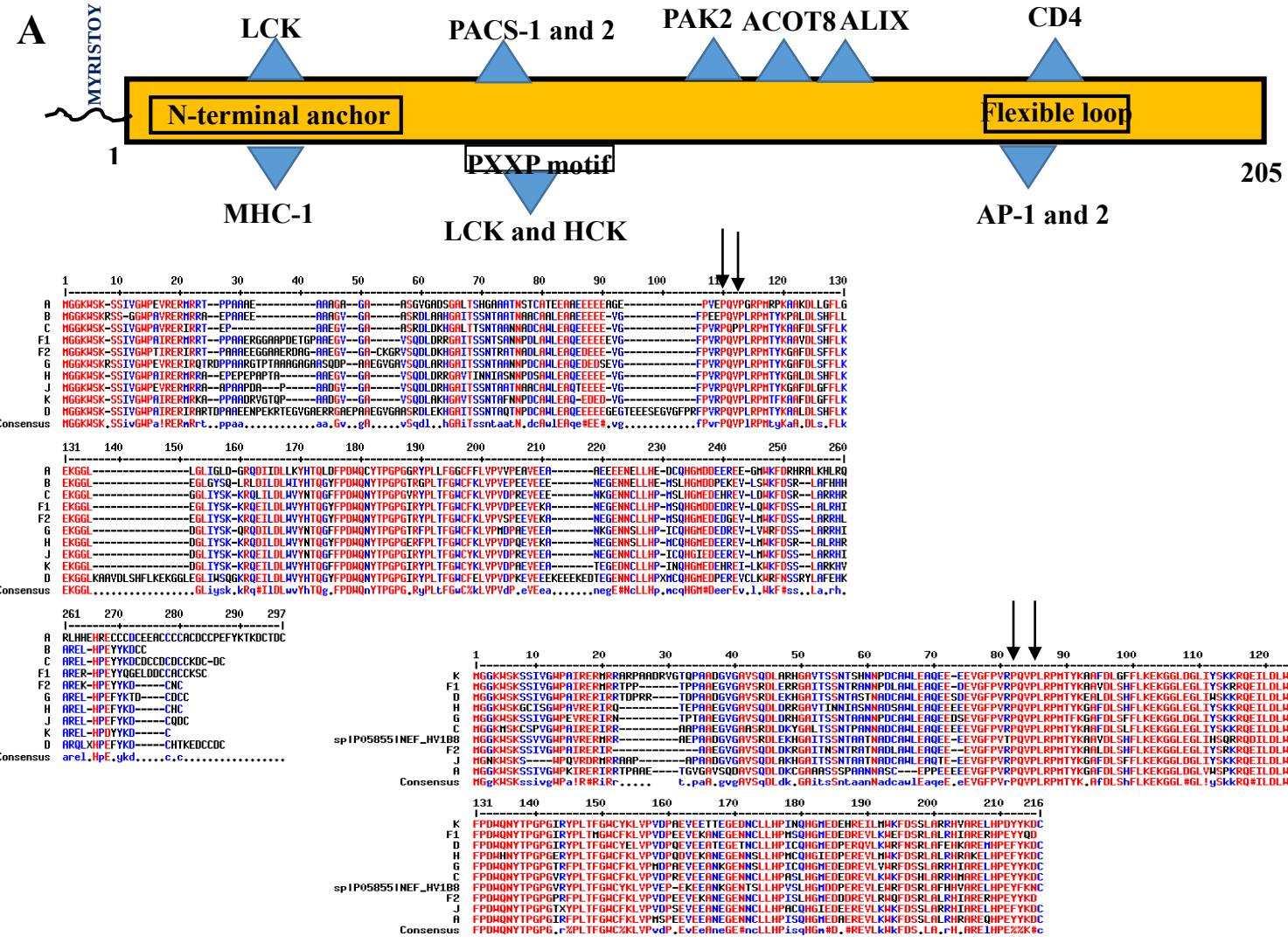
Subtypes		HIV-1 Accessory Proteins			
		Nef	**Vif**	**Vpr**	**Vpu**
A		16	6	4	6
	A1	1659	1408	1375	1830
	A2	61	11	6	5
	A3	8	6	5	5
	A4	3	3	3	3
	A6	224	240	289	282
	A7	3	4	4	4
	A8	2	4	3	3
		1976	1682	1689	2138
B		32288	24983	24765	27341
C		15352	13806	13684	14625
D		762	559	527	742
F	F1	81	164	87	125
	F2	18	60	60	17
G		208	157	159	150
H		23	15	11	10
J		8	7	7	7
K		2	2	2	2

Table 6: Geographical distribution: HIV-1 has been reported from all over the world and stored in the Los Alamos database. According to the reported sequences, subtype A, B and C have been identified in maximum 33, 51 and 41 countries respectively. The least populated subtypes: subtype F2, H, J and K have been reported from 2, 7, 5 and 2 countries respectively. It is quite clear that the countries in the African continent are the most affected in the whole world. However, Subtype B has reported more than 15000 cases in the United State for each accessory protein. The diversity of subtypes is extensive in the African continent but the sequence strength is not so high. One of the reasons behind this may be the lesser identification due to the lower literacy rate. Cameroon and Congo are the top two countries where all the subtypes have been reported. Hence, it can be concluded that African continent possess the highest diversity rate whereas United State contains the highest sequence strength according to the Los Alamos data.

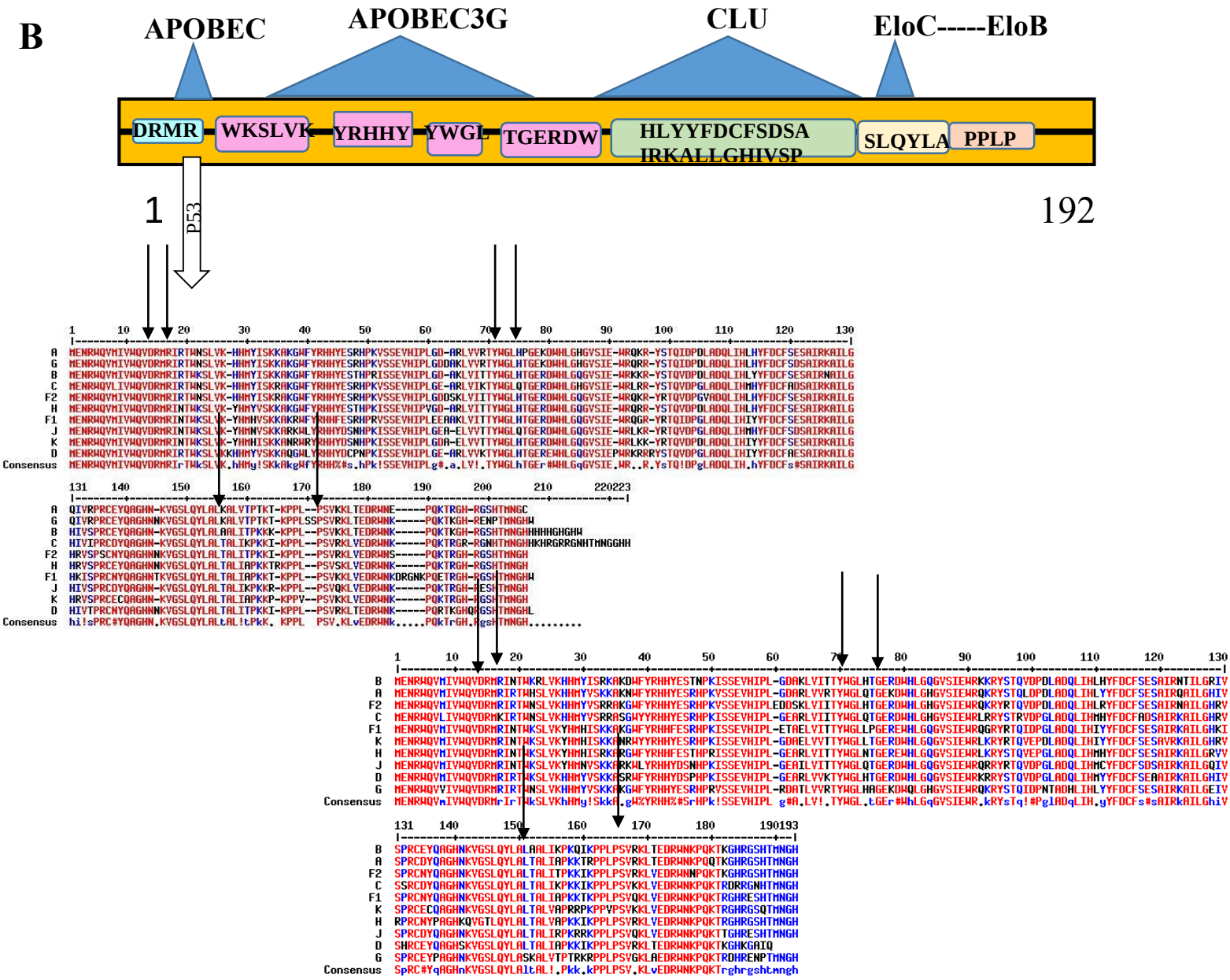
Subtypes	Total number of countries	Top 5 sequence countries	Highest sequence
A	33	Kenya, Pakistan, Russian Federation, Rwanda and Uganda	Kenya
B	51	United State, Canada, South Korea, Australia and Belgium	United State
C	41	Malawi, Tanzania, South Africa, Botswana and Zambia	Malawi
D	22	Cameroon, Kenya, South Korea, Uganda and Congo	Uganda
F1	15	Spain, Brazil, Belgium, Peru, and Argentina	Spain
F2	2	Cameroon and Switzerland	Cameroon
G	24	Cameroon, South Korea, Spain, United State and Nigeria	Cameroon
H	7	Congo, Korea, Congo, United Kingdom and Belgium	Congo
J	5	Cameroon, Angola, Portugal, Congo and Sweden	Congo
K	2	Cameroon and Congo	Equal

Sequence similarity:

Consensus sequence from intra-subtype MSA determined the subtype representatives for four HIV-1 accessory proteins. Multiple insertion/deletion/mutation has been identified in **Vpu** and **Nef** compare to **Vif** and **Vpr**. However, intra-subtype sequence conservancy level is comparable between the subtypes for all 4 proteins. Similar trend has been seen for inter-subtype data also. Amino acids, essential for executing the respective host cell function almost remain conserved among the subtypes. 75% similarity cut-off has been applied to cluster the subtypes with respect to the protein sequences for each accessory protein individually. **Nef** and **Vpu** divided in to 4-7 clusters whereas, **Vif** and **Vpr** proteins show only one cluster each. This further supports the larger sequence diversity in **Nef** and **Vpu** protein across different subtypes. Shannon Entropy data also supports the previous findings having low randomness for **Vif** and **Vpr** proteins (**Figure 2, Table 7 &8**). Nevertheless, Shannon Entropy data also shows that, conservancy in adjacent amino acids with respect to the conserved amnio acids (essential to perform host cell function) are less, which in term can influence the binding of small ligand like molecules.



B


APOBEC
APOBEC3G
CLU
EloC-----EloB
DRMR
WKSLVK
YRHHY
YWGL
TGERDW
HLYYFDCFSDSA
IRKALLGHIVSP
SLQYLA
PPLP
1
192
P53

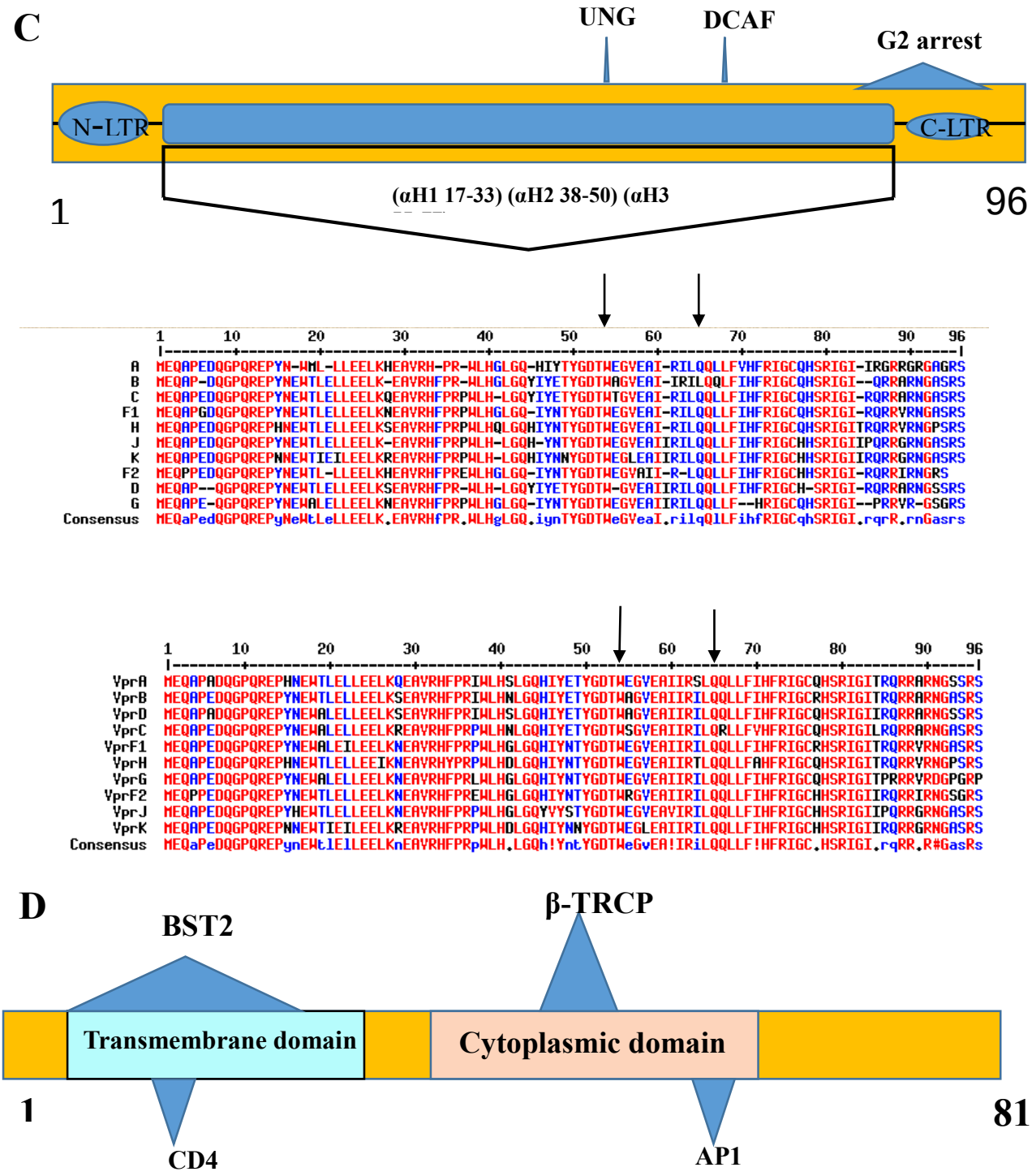

C
UNG
DCAF
G2 arrest
N-LTR
C-LTR
1
(αH1 17-33) (αH2 38-50) (αH3
96
A MEQAPEDQGPQREPYN-WML-LLEELKHEAVRH-PR-WLHGLGQ-HIYTYGDTWEGVEAI-RILQQLLFVHFRIGCQHSRIGI-IRGRRGRGAGRS
B MEQAP-DQGPQREPYNEWTLELLEELK-EAVRHFPR-WLHGLGQYIYETYGDTWAGVEAI-IRILQQLFIHFRIGCQHSRIGI--QRRARNGASRS
C MEQAPEDQGPQREPYNEWTLELLEELKQEAVRHFPRPWLH-LGQYIYETYGDTWTGVEAI-RILQQLLFIHFRIGCQHSRIGI-RQRRARNGASRS
F1 MEQAPGDQGPQREPYNEWTLELLEELKNEAVRHFPRPWLHGLGQ-IYNTYGDTWEGVEAI-RILQQLLFIHFRIGCQHSRIGI-RQRRVRNGASRS
H MEQAPEDQGPQREPHNEWTLELLEELKSEAVRHFPRPWLHQLGQHIYNTYGDTWEGVEAI-RILQQLLFIHFRIGCQHSRIGITRQRRVRNGPSRS
J MEQAPEDQGPQREPYNEWTLELLEELK-EAVRHFPRPWLH-LGQH-YNTYGDTWEGVEAIIRILQQLLFIHFRIGCHHSRIGIIPQRRGRNGASRS
K MEQAPEDQGPQREPNNEWTIEILEELKREAVRHFPRPWLH-LGQHIYNNYGDTWEGLEAIIRILQQLLFIHFRIGCHHSRIGIIRQRRGRNGASRS
F2 MEQPPEDQGPQREPYNEWTL-LLEELKHEAVRHFPREWLHGLGQ-IYNTYGDTWEGVAII-R-LQQLLFIHFRIGCHHSRIGI-RQRRIRNGRS
D MEQAP--QGPQREPYNEWTLELLEELKSEAVRHFPR-WLH-LGQYIYETYGDTW-GVEAIIRILQQLLFIHFRIGCH-SRIGI-RQRRARNGSSRS
G MEQAPE-QGPQREPYNEWALELLEELKNEAVRHFPRPWLHGLGQ-IYNTYGDTWEGVEAIIRILQQLLF--HRIGCQHSRIGI--PRRVR-GSGRS
Consensus MEQaPedQGPQREPyNeWtLeLLEELK.EAVRHfPR.WLHgLGQ.iynTYGDTWeGVeaI.rilqQlLFihfRIGCqhSRIGI.rqrR.rnGasrs
VprA MEQAPADQGPQREPHNEWTLELLEELKQEAVRHFPRIWLHSLGQHIYETYGDTWEGVEAIIRSLQQLLFIHFRIGCQHSRIGITRQRRARNGSSRS
VprB MEQAPEDQGPQREPYNEWTLELLEELKSEAVRHFPRIWLHNLGQHIYETYGDTWAGVEAIIRILQQLLFIHFRIGCRHSRIGITRQRRARNGASRS
VprD MEQAPADQGPQREPYNEWALELLEELKSEAVRHFPRIWLHSLGQHIYETYGDTWAGVEAIIRILQQLLFIHFRIGCQHSRIGIIRQRRARNGSSRS
VprC MEQAPEDQGPQREPYNEWALELLEELKREAVRHFPRPWLHNLGQHIYETYGDTWSGVEAIIRILQRLLFVHFRIGCQHSRIGILRQRRARNGASRS
VprF1 MEQAPEDQGPQREPYNEWALEILEELKNEAVRHFPRPWLHGLGQHIYNTYGDTWEGVEAIIRILQQLLFIHFRIGCRHSRIGITRQRRVRNGASRS
VprH MEQAPEDQGPQREPHNEWTLELLEEIKNEAVRHYPRPWLHDLGQHIYNTYGDTWEGVEAIIRTLQQLLFAHFRIGCQHSRIGITRQRRVRNGPSRS
VprG MEQAPEDQGPQREPYNEWALELLEELKNEAVRHFPRLWLHGLGQHIYNTYGDTWEGVEAIIRILQQLLFIHFRIGCQHSRIGITPRRRVRDGPGRP
VprF2 MEQPPEDQGPQREPYNEWTLELLEELKNEAVRHFPREWLHGLGQHIYNTYGDTWRGVEAIIRILQQLLFIHFRIGCHHSRIGIIRQRRIRNGSGRS
VprJ MEQAPEDQGPQREPYHEWTLELLEELKNEAVRHFPRPWLHGLGQYVYSTYGDTWEGVEAVIRILQQLLFIHFRIGCHHSRIGIIPQRRGRNGASRS
VprK MEQAPEDQGPQREPNNEWTIEILEELKREAVRHFPRPWLHDLGQHIYNNYGDTWEGLEAIIRILQQLLFIHFRIGCHHSRIGIIRQRRGRNGASRS
Consensus MEQaPeDQGPQREPynEWtlEiLEELKnEAVRHFPRpWLH.LGQh!YntYGDTWeGvEA!IRiLQQLLF!HFRIGC.HSRIGI.rqRR.R#GasRs
D
BST2
β-TRCP
Transmembrane domain
Cytoplasmic domain
1
81
CD4
AP1
A MLPPLEICAIVGLIVAL-IIAIVVWTIVGIEYRKLLKQRK----IDRLIERIRERAEDSGNESDGDTEELSTL----V-EMGDYNLGDDNNL
G MQPQSQSLEPLEIAAIVGLIVAF-IAAIIVWTIVFIEYRKIRKQKK----IDRLLDRIRERAEDSGNESEGDTEELATL----V-EMGDFDPWVGDNL
J MIPLQIAAIVALIVAF-FLAIGVWTIVYIEYRKLLRQRK----IDKLIDRIRERAEDSGNESDGDTEELAKL----V-EMGPHDLWNVNDL
B MQPLQTLPLQILAIVALVVAA-IIAIVVWTIVFIEYRKILRQRK----IDRLIDRIRERAEDSGNESEGDQEELSAL----V-EMGHLAPWDVDDLLHHDDDDDDDLDDDDDLLDLLDLLDL
C MLDLLARVDYRLGVGALIVAL-IIAIVVWIIAYIEYRKLLRQRK----IDWLIKRIRERAEDSGNESEGDTEELSIM----V-DMGHLRLLDVNDLLDLDDDLLLEEGDKCGDRLLDVNDL
H MNIMYLGILGIGALIVAF-IIAIIVWTIVYIEYRKLVKQKK----IDRLIERIRERAEDSGNESDGDTEELSTL----V-EMGHLNLGYVADL
F1 MSDMLLIISIVALIVAL-IIVIVVWTIVYIEYRKLVRQRK----INRLYERIRERAEDSGNESEGDAEELAAL----GGEVGPFIPGDINNL
F2 MSYISILIILVIVAFIVAL-IIAIIVWTIVYIEYKKLLRQKR----INRLYERIRERAEDSGNESEGDAEELAAL----G-EVGPFIPGDINNLC
K MVPLAISGIIALVAAL-ILAIIVWTIAYIEYRKLVKQKR----INWLFDRIRERAEDSGNESEGDAEELAAI----G-ELGHLILGDIDNL
D MQPLVILAIVALVVALXILAIVVWTIVFIECRKRLRRQRKNWLIDWLIDRIRERAEDSGNESEGDREENKKLSAANLVEMGHHAPVEMGHHAHHDGYHDAAPWDVDDMLNIK
Consensus .....pL.I.aIvaL!vAl.iiAI.VWTIvyIEyRKllrq.k....I#rLi#RIRERAEDSGNES#GD.EEla.l....v.EmGh..pgd..nl.....
B MQPLQILAIVALVVAAIIAIVVWTIVIEYRKILRQRKIDRLIDRIRERAEDSGNESEGDQVELSALVEMGHHAPWDVDDLL
A MQPLVILAIVALVVTLIIAIVVWTIVFIEIRKIRRQRKIDRLIDRIRERAEDSGNESEGDTEELSKLVEMGHDAPWDVDDL
D MQPLGIIAIAALVVAIILAIVVWTIVFIEYRRIKKQRRIDCLLDRITERAEDSGNESEGDREKLSKLVEMGHHAPWDIDDL
F1 MSNLLAIGIAALIVALIITIVVWTIAYIEYKKLVRQRKINRLYKRISERAEDSGNESEGDAEELAALGEVGPFIPGDINNL
C MLELIGRIDYRLGVGALIVALIIVIIVWTIAYIEYRKLVRQRRIDWLVKRIKERAEDSGNESGGDTEELETMVDMGHLRLLDGNDL
H MYILGLGIGALVVTFIIAVIVWTIVYIEYKKLVRQKKIDRLIERIGERAEDSGNESDGDTEELSKLMEMGHLNLGYVADL
Consensus .....lg!gAL!ValIIa!!VWTIv%IEYrKlvRQrkI#rL..RI.ERAEDSGNES.GDtEELs.$v#MGhl.l.d..#L

Figure 2 (A-D): Inter-subtype MSA of four HIV-1 accessory proteins. The important regions for protein-protein interactions are denoted with the arrows. It is shown that the insertion opoint mutation is higher in case of **Vpu** and **Nef**. In most of the cases the sequence difference has been spotted out on both terminal domains. N-terminal domain of HIV-1 Nef protein is an IDPr and has the highest sequence variation whereas sequence variation is present on the C-terminal domain of **Vif** and **Vpr**. Here, both the data from LANL database and UniProt database has been shown. Amino Acids which play essential role in host cell functions are shown to be conserved across the subtypes. Though the conservancy level in PPI region is lesser in case of **Vpu** where other three proteins have comparable data. However, mutations are spotted in amino acid residues adjacent to the important amino acids for each accessory proteins.

Table 7: Clustering analysis of Protein sequences. Subtype B and C contain more than 12000 sequences for each accessory protein. Hence, Intra-subtype clustering with 75% similarity cut-off help to determine the distinct sequences for each protein. The distinct sequences are shown to be from the countries where the number of reported sequences is high. Data reveals that **Vif** and **Vpr** has the least sequence diversity among the four accessory proteins while Nef and Vpu protein have a comparatively lesser sequence conservation in a subtype. Inter-subtype clustering of this accessory proteins with 75% similarity cut-off also suggests that **Vpr** and **Vif** has less variation across the subtypes with predominant subtype B or C as representative while **Nef** and **Vpu** have more variation among the subtypes.

Alignment	Protein	Subtype	Sequence strength	Cluster identified	Cluster representatives
Intra	**Nef**	B	47	6	FR-16 US-22 CA-1 BE-6 AU-1 BR-1
		C	41	4	IN-1 ZA-31 ZM-8 RU-1
	Vif	B	50	1	US-50
		C	33	1	MW-33
	Vpu	B	51	3	BR-40 SE-1 PE-10
		C	38	5	ET-18 BW-4 IN-3 ZA-10 UY-1
	Vpr	B	50	1	US-50
		C	37	1	ZA-37
Inter	**Nef**	A, B, C, D, F1, F2, G, H, J and K subtypes	10	4	D A G- F1, F2, G, H and J C- B, C and K
	Vif	Same	10	1	C- All
	Vpu	Same	10	7	B C D G – G, F1 and F2 H - A and H J K
	Vpr	Same	10	1	B- All

Table 8: Amino acids or motifs of the accessory proteins of HIV-1 which are essential for the protein-protein interaction between host and virus. Similar types of mutation have been seen for each accessory protein in the sequences from both databases. However, this data also suggest that the virus maintains the essential amino acids throughout the subtype to execute the cellular functions.

Accessory Proteins	Important amino acids or motifs	Subtypes	Mutation	
			Los Alamos sequence	UniProt sequence
Nef	PxxP motif-SH3 binding	-	-	-
Vif	DRMR-A3F binding	C	-	M17K
	YRHHY-A3G	-	-	-
	TGERDW-A3G	A and G	R77K	R77K
		A	T74P	
		G		T74A
		F1		T74P
	SLQYLA-EloC EloB	H	-	S144H
	PPLP	K	L163V	L163V
Vpr	LLeeL-NLS-1	K	L23I	L23I
		F1	-	L23I
		H	-	L27
	W-UNG2	-	-	-
	LQqLLxxxxRIgcR-DCAF-1	B	One deletion of L67	-
		C	-	Q66R
	RQRRxR-G2 arrest	B	Deletion of R85	-
		G		R85P and
		J	R85P	Q86R
				R85P
Vpu	ALxxAxxxAxxxW-BST2	A and G	A10G	-
		F2	L11F	-
		F1	A18V	A18T
		C	-	A18V
	ExxxxxLV-βTRCp	C	L to M	L to M

Phylogenetic analysis

This study is conducted to infer the evolutionary relationships between the subtypes corresponding to the accessory protein sequences. The data is not capable of telling the full scene of divergence. However, 2 to 3 facts are generated by this analysis **(Figure 3).** For Nef protein, 3 different clusters (D, G, H and K), (F1 and F2) and (A, B, C and J) are formed with a higher distance level. Whereas for the other three protein the distance is relatively low among the subtypes. F1 and F2 are seem to be near neighbour except Vpr protein. Similarly, it can be speculated that sub type B and C diverged from common ancestor. However, diversity has been seen in the dendograms of the proteins which indicates differences in the evolutionary and functional aspects.

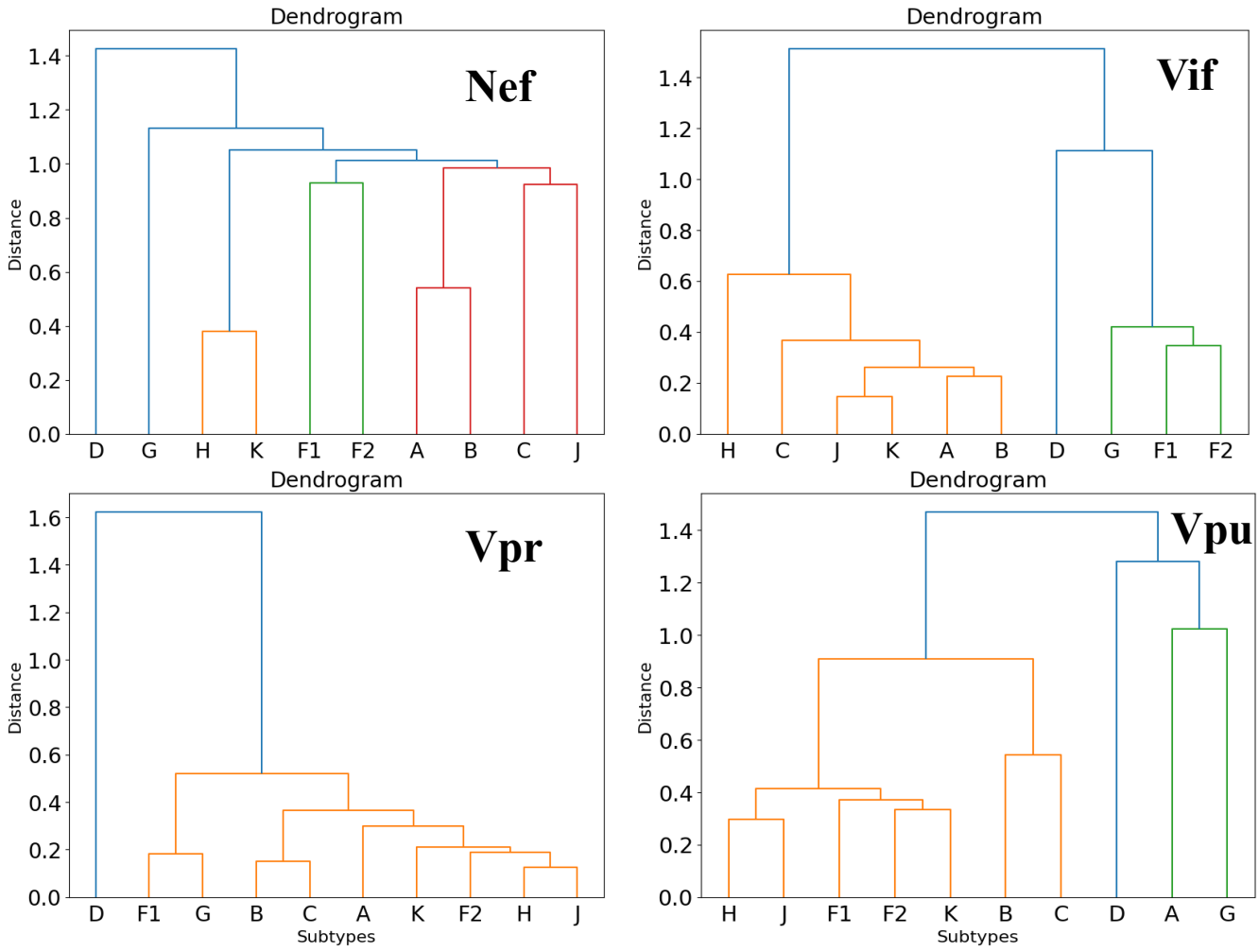
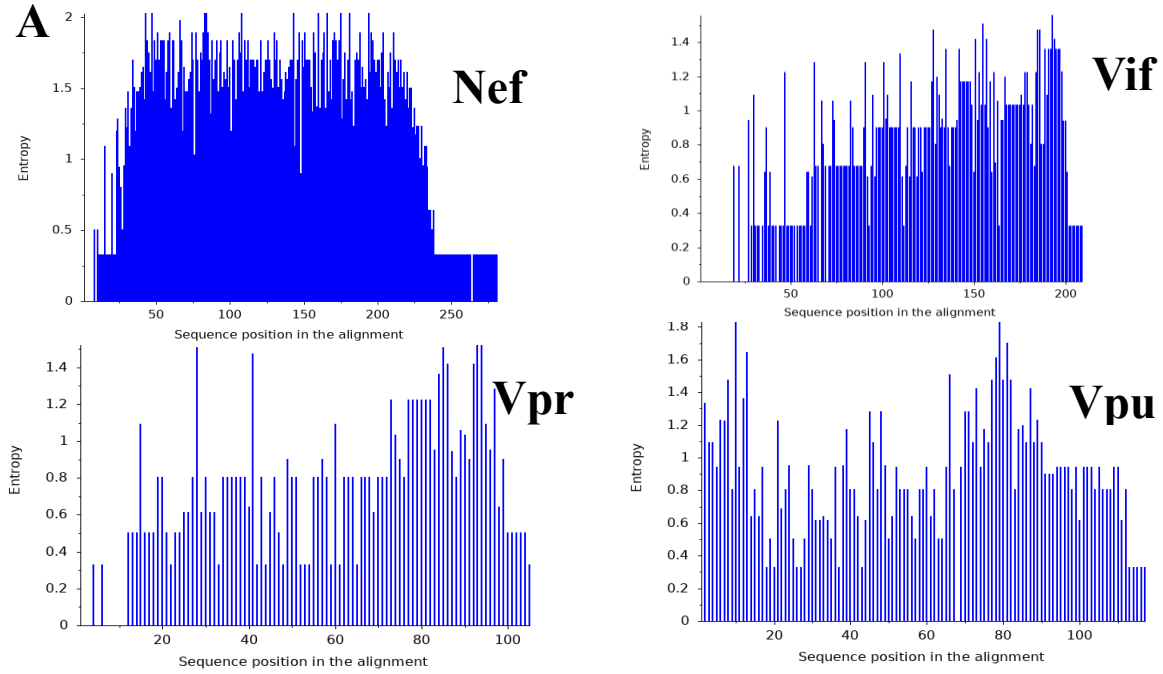


Figure 3: This analysis is executed to get the idea of the phylogenetic relation between the subtypes. Considering ten (A, B, C, D, F1, F2, G, H, J and K) subtypes, individual dendogram has been plotted for all accessory proteins. Refer to these dendograms, it can be concluded that subtype D has diverged in each case at a higher distance. Subtype G could be predicted as the

near neighbour of subtype D. F1 and F2 with shortest distance apart in each case supports that they are diverged from one single subtype F. It can be said that subtype B and C diverge from one single ancestor as they are shown to be closely related at a very lower distance. An assumption can be made that F1 and F2 cluster is closely related to the cluster G as in most of the cases they eventually merged to the subtype G cluster.

Determination of Shannon Entropy

Shannon entropy data also supports the finding that **Nef** and **Vpu** have higher randomness i.e. sequence variability whereas **Vpr** and **Vif** are more conserved across the subtypes. Rough estimation can be interpreted from the consolidated subtype data of the accessory protein individually that subtype B is the most conserved among the subtypes **(Figure 4).** An uneven entropy distribution has been seen in general. Where the Nef protein shows a very high entropy, Vif protein shows the least entropy distribution. The region 50-80 shows the highest entropy in Vpu protein than the terminal positions whereas, both the terminals have the highest entropy in Vpr protein. This may influence the sequence diversity and in the functional properties of the accessory proteins. When it comes among the subtypes, subtype B (index 2) has the lowest entropy indicating the least sequence variability in the subtypes.



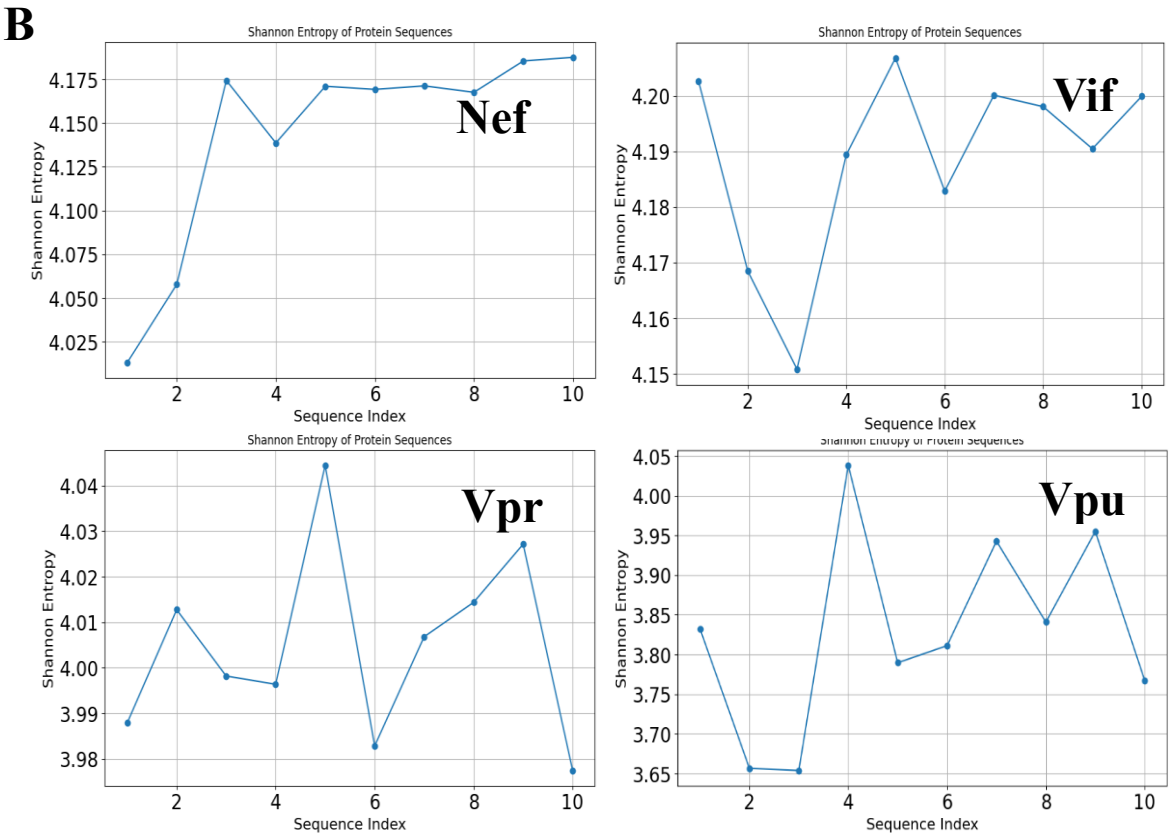


Figure 4: The extent of sequence conservation in accessory proteins has been determined by Shannon Entropy. (A). interprets the sequence conservation in each position of an accessory protein considering all 10 subtypes. It is clearly seen that **Nef** and **Vpu** have higher Shannon entropy value than the other two accessory proteins. This supports the previous findings of both proteins having the highest sequence variation. (B). the consolidated Shannon entropy of individual subtypes for each accessory protein. Subtype B satisfies to be the most conserved except for **Vpr** protein. Subtype K shows low Shannon entropy considering the **Vpr** and **Vpu** protein but reaches the highest for **Nef** and **Vif** protein. While subtype C is also showing to have least variation with respect to **Vif ,Vpr** and **Vpu** proteins, but almost reaches above value 4 considering **Nef** protein sequences. However, the sequence strength of subtype H, J and K are almost negligible comparing to subtype A, B, C and D. In conclusion it can be said that subtype B and C maintain the sequence variation in least extent than other subtypes though it is complicated to draw any conclusion of all the subtypes from the data. The amino acids which are important for each accessory protein.

Structural similarity

Secondary structures of the proteins are almost similar for 10 subtypes for each accessory protein. Though amino acid in one position may vary from subtype to subtype, but the virus maintains the identical structure of a particular stretch. Here, the secondary structures of **Nef** proteins from all subtypes have been displayed as it is the largest accessory protein and has the highest sequence variation among the subtypes. Subtype F2 has some deletion in the protein sequence whereas subtype D and K have insertion of few amino acids. It is cleared from that in spite of having sequence variation like insertion, deletion or any point mutation, the motifs maintain their structure across the subtypes. Similar result is shown for each accessory protein (**Figure 5**).

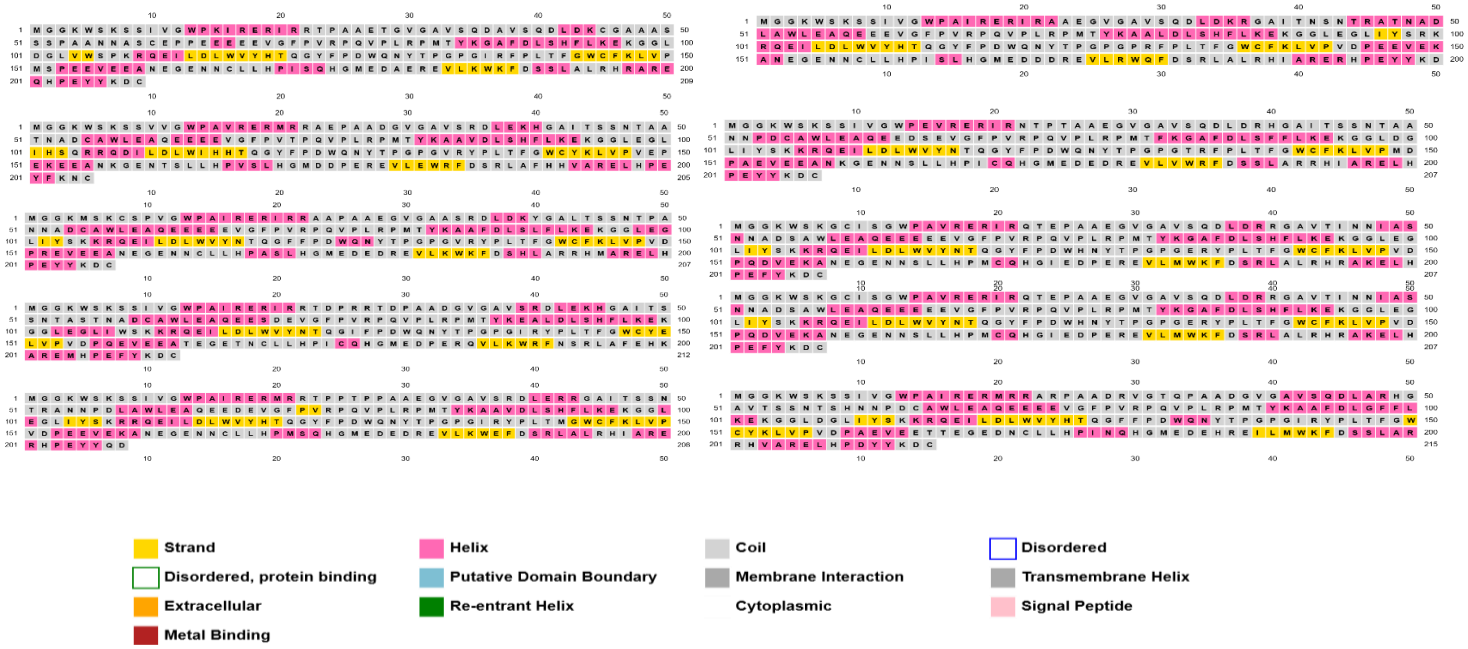


Figure 5: The secondary structure of all 10 subtypes of HIV-1 **Nef** protein. The left panel shows subtype A-F1 and right panel shows subtype F2-K (from top to bottom). Similar work has been executed for other three accessory proteins. Secondary structure similarity has been found out for all the proteins across the subtypes.

Tertiary structures were modelled for all the accessory proteins for all 10 subtypes using the representative consensus sequences using the PDB template. The protein structures from **Vif** and **Vpr** proteins are shown to be well superimposed with each other. Few minor differences can be shown in the terminal domain for the **Vpu** protein having well superimposed structure in the coil region. However, a lot more differences can be seen between the superimposed structures of **Nef** protein (**Figure 6 & 7**).

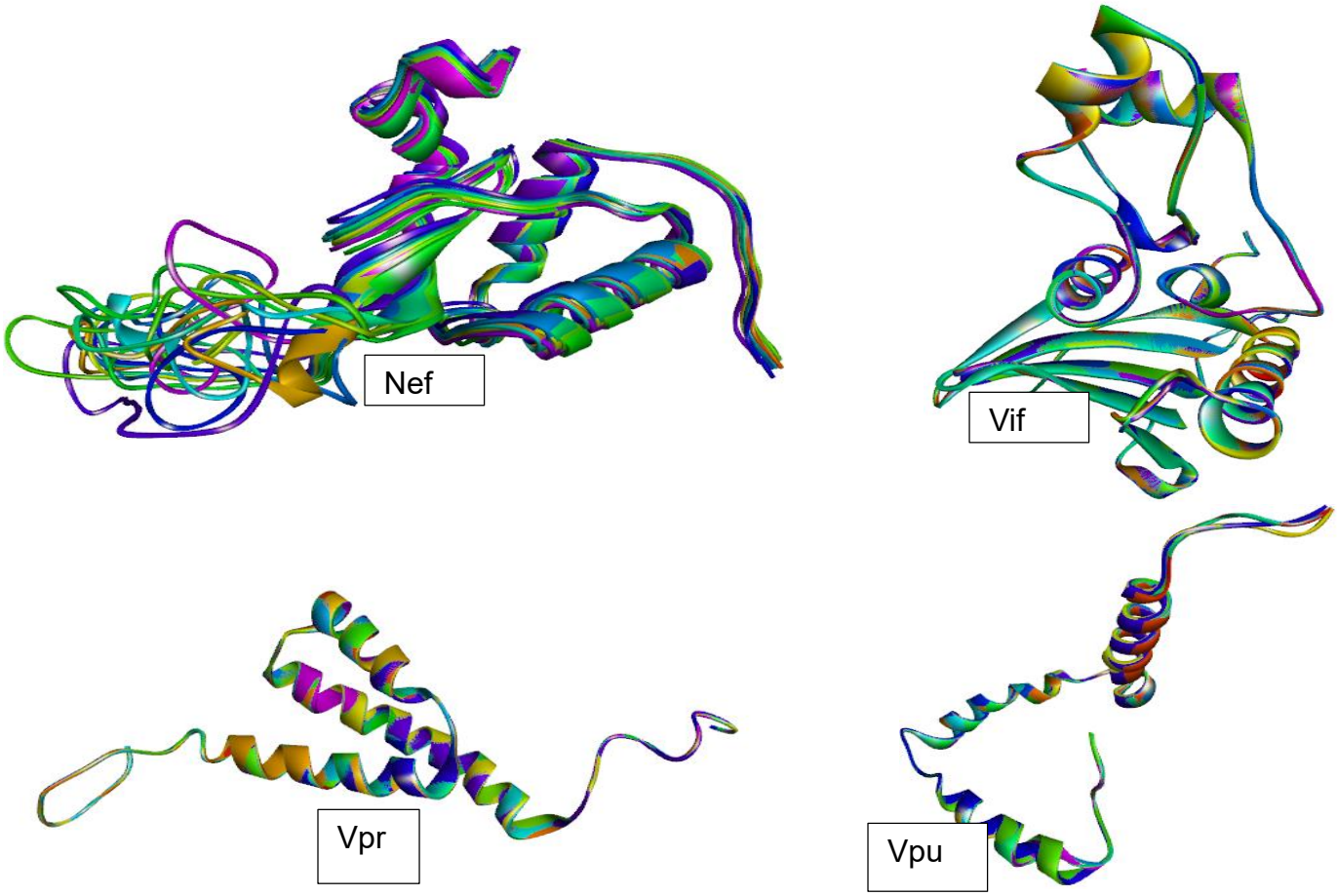


Figure 6: Superimposed structures of the accessory proteins.

	NefA	NefB	NefC	NefD	NefF1	NefF2	NefG	NefH	NefJ	NefK
NefA	0									
NefB	3.296	0								
NefC	3.273	3.756	0							
NefD	3.992	4.459	2.684	0						
NefF1	3.632	3.951	3.298	3.463	0					
NefF2	2.314	4.158	3.972	3.981	4.44	0				
NefG	2.736	4.007	2.34	3.09	3.918	3.73	0			
NefH	2.86	4.445	3.046	3.505	4.532	3.11	3.154	0		
NefJ	3.398	4.523	2.681	3.409	3.611	3.861	2.48	3.371	0	
NefK	4.166	5.93	4.355	4.13	4.702	5.272	3.441	4.687	3.374	0

	VifA	VifB	VifC	VifD	VifF1	VifF2	VifG	VifH	VifJ	VifK
VifA	0									
VifB	0.038	0								
VifC	0.041	0.045	0							
VifD	0.042	0.043	0.049	0						
VifF1	0.054	0.053	0.053	0.058	0					
VifF2	0.147	0.147	0.15	0.147	0.146	0				
VifG	0.051	0.057	0.058	0.058	0.061	0.014	0			
VifH	0.052	0.052	0.044	0.053	0.055	0.033	0.063	0		
VifJ	0.051	0.048	0.051	0.043	0.054	0.02	0.064	0.054	0	
VifK	0.045	0.041	0.049	0.049	0.046	0.057	0.061	0.05	0.046	0

	VPRA	VPRB	VPRC	VPRD	VPRF1	VPRF2	VPRG	VPRH	VPRJ	VPRK
VPRA	0									
VPRB	0.16	0								
VPRC	0.24	0.26	0							
VPRD	0.04	0.15	0.24	0						
VPRF1	0.17	0.07	0.26	0.16	0					
VPRF2	0.19	0.08	0.33	0.19	0.18	0				
VPRG	0.34	0.37	0.38	0.34	0.35	0.38	0			
VPRH	0.22	0.28	0.29	0.24	0.28	0.34	0.37	0		
VPRJ	0.34	0.29	0.42	0.33	0.28	0.3	0.37	0.4	0	
VPRK	0.22	0.07	0.3	0.18	0.08	0.15	0.37	0.3	0.29	0

	VpuA	VpuB	VpuC	VpuD	VpuF1	VpuH
VpuA	0					
VpuB	1.006	0				
VpuC	0.334	0.834	0			
VpuD	0.073	1	0.325	0		
VpuF1	0.628	0.713	0.75	0.632	0	
VpuH	0.8	0.822	0.36	0.793	0.842	0

Figure 6 and 7: The superimposed tertiary structures of HIV-1 accessory proteins. **Nef** (A): Template pdb 4U5w.1.D (71-207: excluding the disordered N terminal domain). The structures have a higher difference in the loop region and a small difference in the coil region. According to the inter-subtype MSA discussed in the previously, it is expected that with the higher sequence variation the structure of **Nef** protein can vary from subtype to subtype. However, the virus maintains the same secondary structure across the subtypes but they differ with a high RMSD value between the structures. **Vif** (B): Template pdb 8CX.0.1.B (1-176: excluding the C terminal domain). The structures are well superimposed with each other. Inter-subtype RMSD doesn't exist 0.15Å which is much less than conventional cut-off value. Supporting the findings from sequence level data it has been again seen that **Vif** protein has the least variation and RMSD across the subtypes. **Vpr** (C): Template pdb: 1M8L.1.A. The structures are well

superimposed with highest RMSD of 0.4Å. **Vpu** (D): Template pdb: 2n28.1.A. The structures are well superimposed in the coil region but has a difference in the terminal domains. The highest RMSD is near 1.0Å proving the sequence variation between the subtypes.

Structure validation

Ramachandran Plot satisfies the modelled structures for all subtypes of 10 subtypes of HIV-1. The dynamic properties of the **Vpr** protein models were determined through Molecular Dynamics Simulation in different force field and parameters (specifies in the method section) (**Figure 8)**. Along with that lowest energy conformation of **Vpr** protein from each subtype were identified (**Figure 9**). **Nef** protein structures were again validated using MD simulation with appropriate force field (**Figure 10**). Different conformations from clustering of the MD trajectories have been listed in **Table 9**.

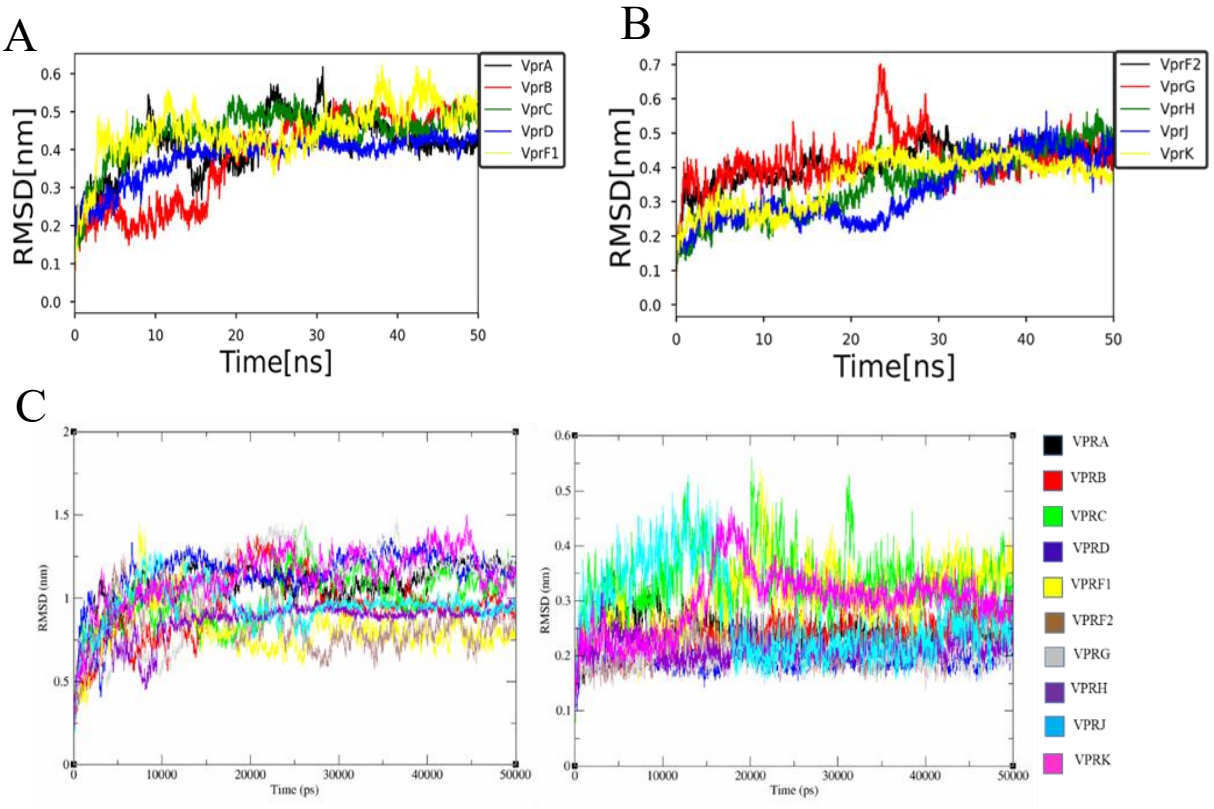
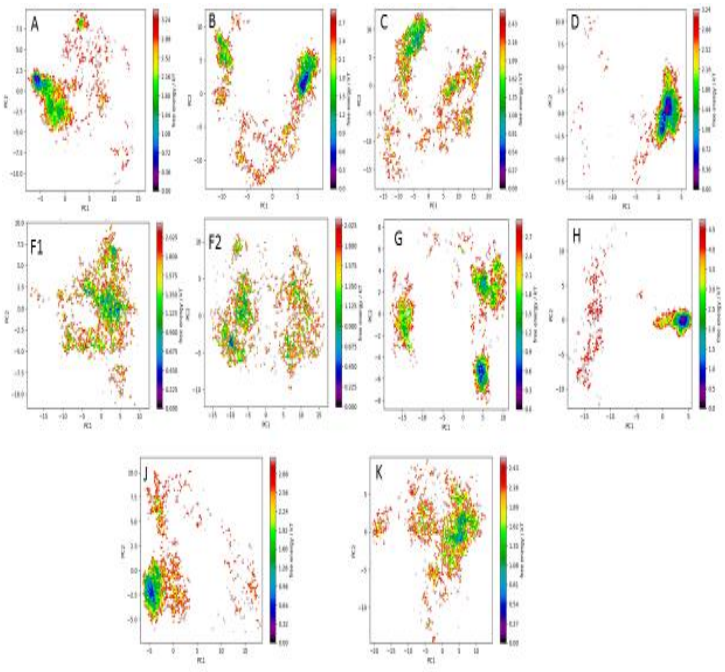


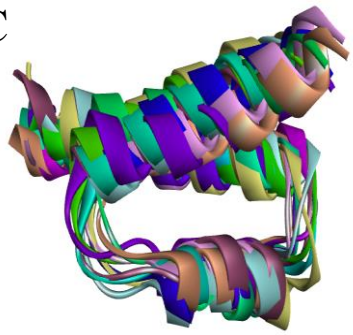
Figure 8: RMSD plot of **Vpr** protein. (A and B) RMSD of **Vpr** helix (17-77) from subtype A-F1 and F2-K respectively in Amber Force Field. (C) RMSD of full (left) and truncated (**Vpr** helix:17-77) **Vpr** in Charmm36 Force Field. It is predetermined that **Vpr** proteins from 10

subtypes are well superimposed with each other. The dynamic nature of the protein also supports that all the protein models reach a plateau after 35ns (full trajectory=50ns) in each case irrespective of the Force Field. However, the full **Vpr** protein has shown a higher RMSD of 1.0Å than the truncated one where the RMSD lies in between 0.2-0.4Å. The loop structure of both the terminal domains is responsible for the higher deviation. RMSF result also support the findings having higher fluctuation in the terminal residues than the helix region for each subtype.

A



C



B

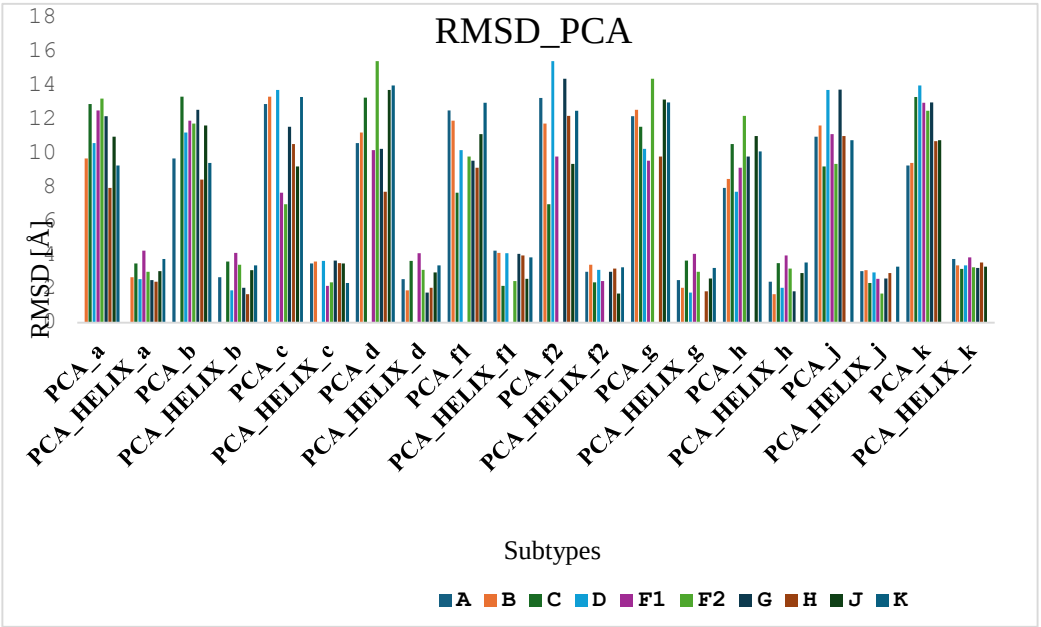


Figure 9: Lowest free energy conformation of **Vpr** proteins by PCA followed by free energy surface analysis. A: Clustering of the individual subtypes to identify the free energy conformation. B: The RMSD of the lowest energy structures of all the subtypes with respect to one subtype individually. C: Superimposition of representatives from the lowest energy conformation. However, The RMSD of the helix portion varies within 2-3Å for each subtype whereas, the full **Vpr** protein conformations have very higher RMSD. Hence, only helix regions are superimposed which shows the differences in the lowest energy conformations among the subtypes. In conclusion it can be said that the subtypes may differ in conformation in spite of having similarity from sequence to structure level.

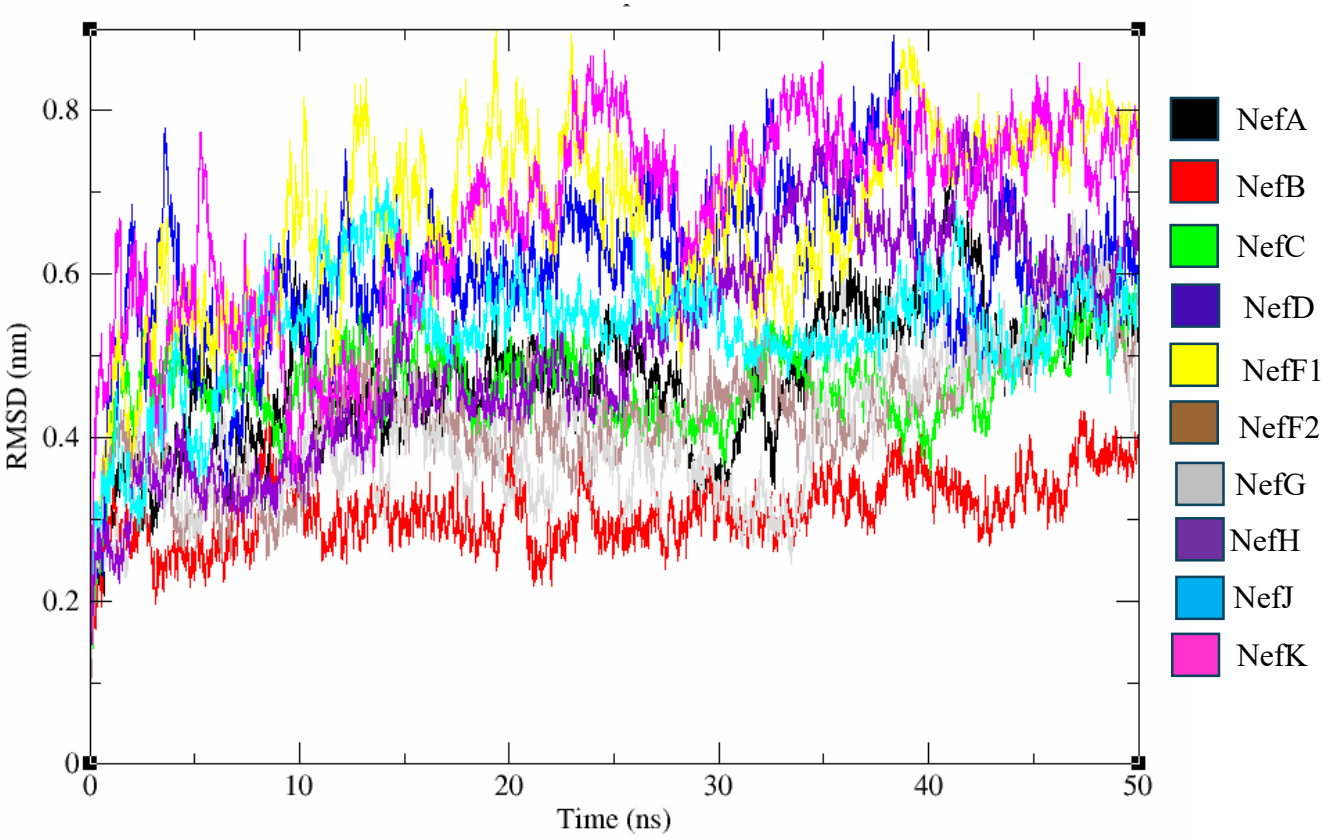


Figure 10: Inter-subtype RMSD of Nef protein (71-207: excluding the anchor domain). The plateau for each subtype was almost reached after 40ns of trajectory of MD simulation in Charmm27 Force Field. The RMSD of the model structure varies within a wide range of 0.2-0.8 Å supporting the highest sequence variation among the subtypes. In the Raw structure superimposition (Figure 6), it was seen that Nef protein from subtypes F1 and K varied higher than other subtypes. Here also these 2 subtypes have higher RMSD compared to the other subtypes.

Table 9: Cluster representatives of Nef proteins extracted from individual MD simulation trajectory. It is clear that subtype B makes an exception of having very small number of cluster from all the three trajectories. Subtype B and subtype C have the lowest number of total cluster representatives by having 4 and 6 cluster representatives respectively. Nef protein has the highest sequence variation across the subtypes with differences in tertiary structure also. Here, it can be said that Nef protein can have multiple conformations varying from subtype to subtype.

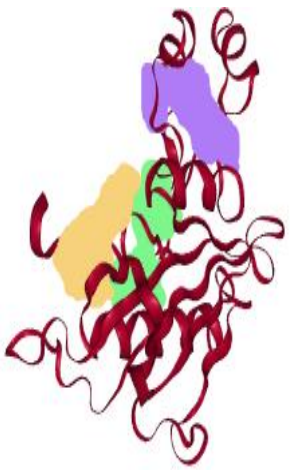
Subtypes	Simulation replicates (triplicate for each subtype 50ns	Total number of clusters	Number of cluster representatives
A	A1	14	2: A1_1, A1_2
	A2	33	3: A2_1, A2_2, A2_3
	A3	48	3: A3_1, A3_2, A3_3
B	B1	6	1: B1_1
	B2	4	1: B2_1
	B3	6	2: B3_1, B3_2
C	C1	22	3: C1_1, C1_2, C1_3
	C2	23	2: C2_1, C2_2
	C3	29	1: C3_1
D	D1	23	2: D1_1, D1_2
	D2	42	3: D2_1, D2_2, D2_3
	D3	26	3: D3_1, D3_2, D3_3
F1	F11	31	2: F11_1, F11_2
	F12	25	3: F12_1, F12_2, F12_3
	F13	17	3: F13_1, F13_2, F13_3
F2	F21	18	4: F21_1, F21_2, F21_3, F21_4
	F22	6	2: F22_1, F22_2
	F23	12	2: F23_1, F23_2
G	G1	17	2: G1_1, G1_2
	G2	23	4: G2_1, G2_2, G2_3, G2_4
	G3	22	2: G3_1, G3_2
H	H1	17	3: H1_1, H1_2, H1_3
	H2	23	3: H2_1, H2_2, h2_3
	H3	24	3: H3_1, H3_2, H3_3
J	J1	22	2: J1_1, J1_2
	J2	14	3: J2_1, J2_2, J2_3
	J3	27	3: J3_1, J3_2, J3_3
K	K1	27	2: K1_1, K1_2
	K2	29	3: K2_1, K2_2, K2_3
	K3	32	3: K3_1, K3_2, K3_3

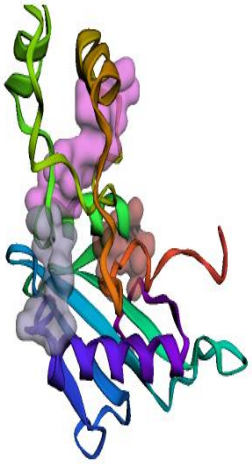
OBJECTIVE 2

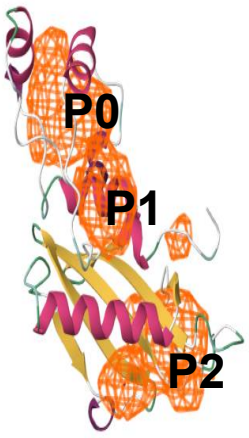
This objective mainly deals with the druggable pocket of the accessory proteins of HIV-1. It is very important to get elaborative information about the druggable pockets of the protein across the 10 prominent subtypes. Characterization of the druggable pockets with respect to different geometrical and physicochemical aspects may provide some idea in futuristic drug discovery study. Here, all of the four accessory proteins of HIV-1 were studied where Nef protein has been studied with a different approach than others.

Vif- The potential druggable pocket of HIV-1 Vif protein of all the subtypes has been predicted with 3 consecutive webservers to make the search more stringent. Here, the pocket forming residues are identified and listed according to the data predicted by web servers: Dogsitescorer, Fpocket and Deepsite. Though the scoring systems are different for each server, it is quite clear that the main 3 pockets constitute of the similar amino acid residues throughout the Vif protein of 10 HIV-1 subtypes (**Figure 11**). The region with the highest probe binding capacity is seen to be identical for all the sub types. The least sequence diversity of Vif protein may be the most probable reason for this.

A

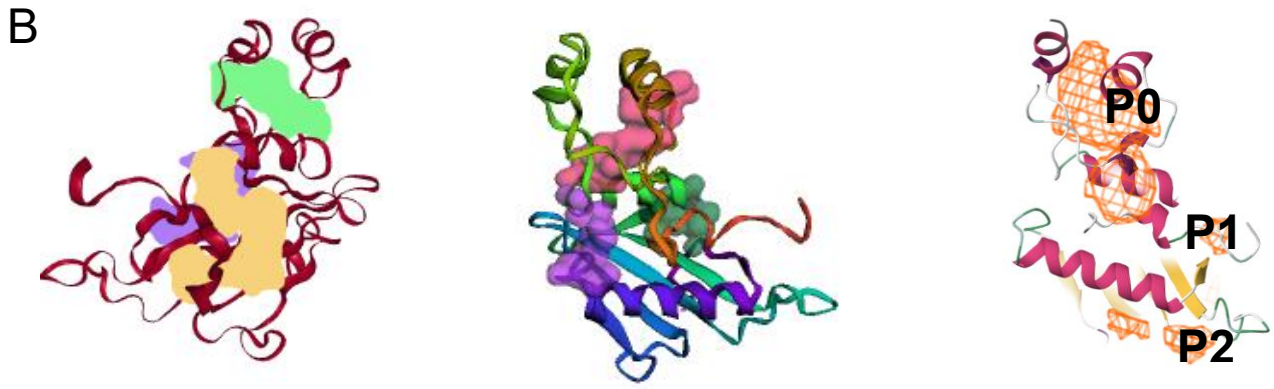






Dogsite Scorer				
Pocket name	Volume Å^3	Surface Å^2	Drug Score	Simple Score
P_0	379.33	690.89	0.7	0.2
P_1	363.39	674.47	0.67	0.23
P_2	356.22	528.09	0.75	0.24

Fpocket Web				
Pocket name	Score	Druggability	Alpha Spheres	Volume
P_0	0.232	0.003	29	322.766
P_1	0.202	0.955	52	727.679
P_2	0.182	0.002	23	143.373



Dogsite Scorer				
Pocket name	Volume Å³	Surface Å²	Drug Score	Simple Score
P_0	434.75	701.61	0.67	0.32
P_1	390.34	587.2	0.72	0.26
P_2	340.67	680.9	0.66	0.22

Fpocket Web				
Pocket name	Score	Druggability	Alpha Spheres	Volume
P_0	0.209	0.008	27	144.022
P_1	0.197	0.947	54	739.238
P_2	0.154	0.001	17	313.680

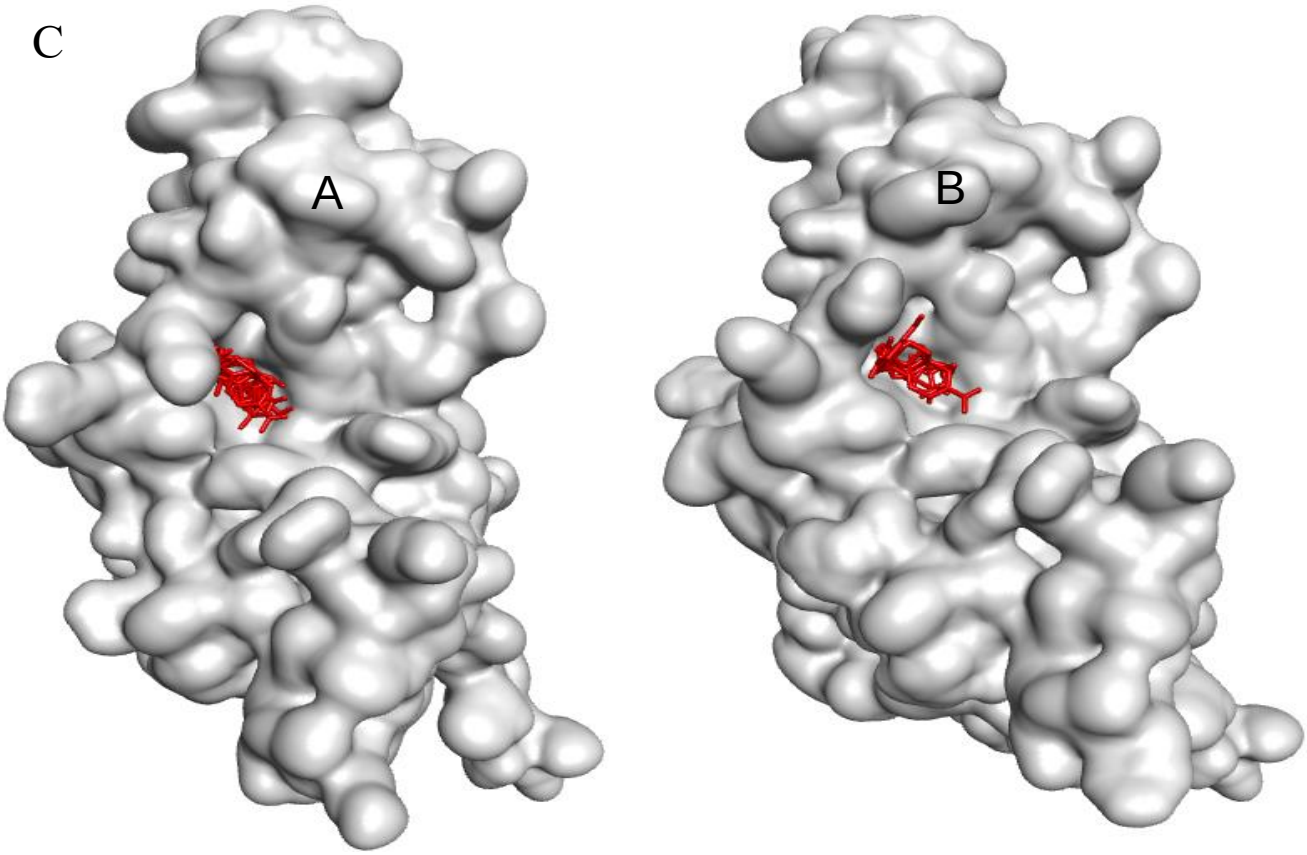
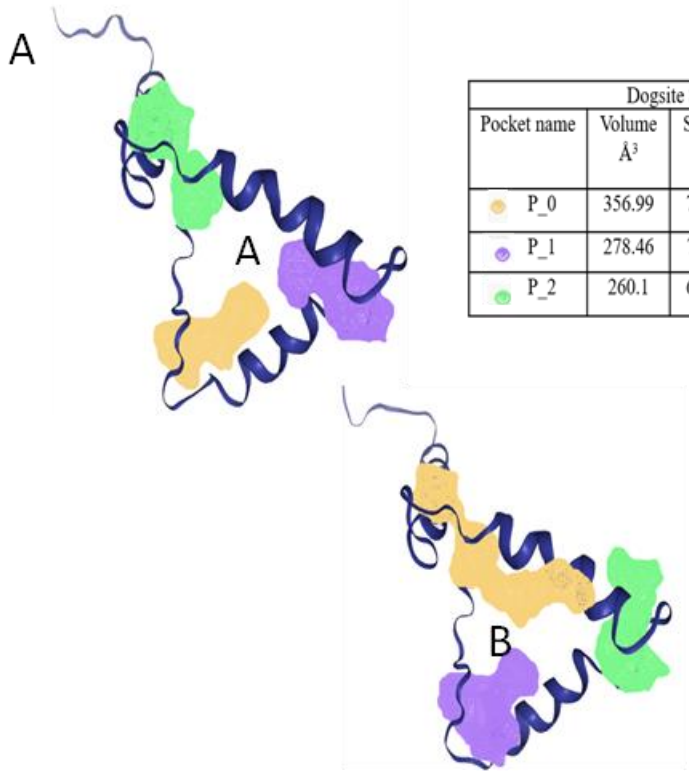


Figure 11: (A) Pockets of Vif protein from subtype A, (B) Pockets of Vif protein from subtype B, (C) Common region of highest probe binding cluster. (Rest of the subtypes are not shown

here). It has been shown that one specific pocket differs in geometrical properties for different subtypes as the ranking varies from subtype to subtype. However, while analysing the FtMap result, it is shown that the highest cluster of probes binds in the similar region for all the subtypes. Hence, from all the result it can be concluded that difference can be there for small molecule binding in a particular pocket between the subtypes.

Vpu- Aforementioned web servers are also used for predicting the pockets for Vpu protein. Here also we have seen that primary three pockets are similar across the subtypes but their ranking differs from subtype to subtype. Unlike Vif protein, Vpu protein also face differences in FtMap result also. For each subtype the highest probe binding cluster become at different region of protein. Thus, having close sequence and structural similarity the subtypes have difference in pocket geometry and probability to bind with several small probe-like molecules (**Figure 12**). Interestingly, the common pocket in the helix region for subtype A and subtype B differs in ranking however, the properties of the pockets are seen to be comparable. In FtMap result, it is clear that the highest probe binding cluster lies on the helix area for 6 subtypes but the regions are not identical.



Dogsite Scorer				
Pocket name	Volume Å³	Surface Å²	Drug Score	Simple Score
P_0	356.99	777.55	0.51	0.22
P_1	278.46	739.17	0.43	0.15
P_2	260.1	688.84	0.45	0.15

Dogsite Scorer				
Pocket name	Volume Å³	Surface Å²	Drug Score	Simple Score
P_0	427.33	1304.77	0.87	0.32
P_1	400.26	941.56	0.51	0.32
P_2	254.53	942.79	0.34	0.13

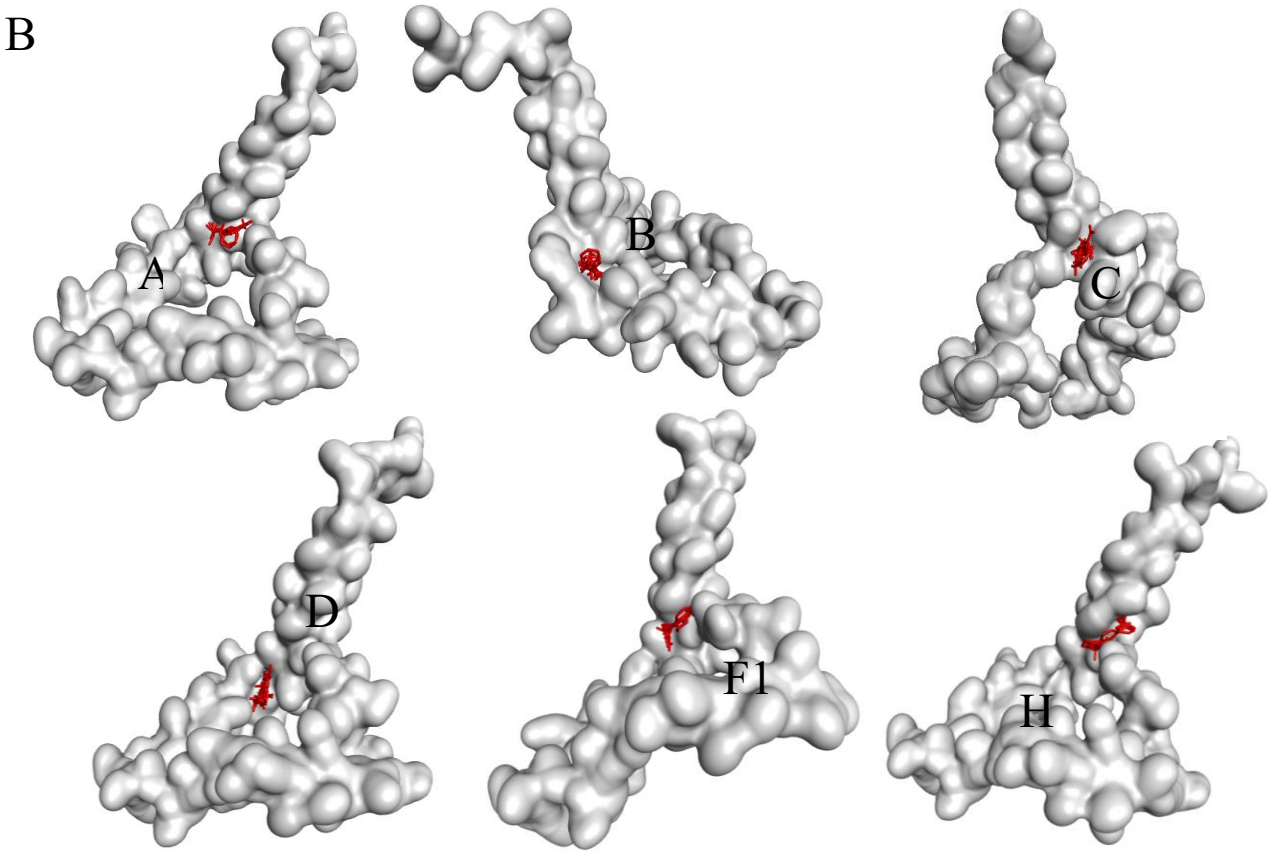


Figure 12: (A) Pockets of Vpu protein from subtype A and B (rest of the subtypes are not shown here). (B) Highest probe binding cluster of Vpu protein from 6 subtypes. The pockets differ in the geometry from subtype to subtype. FtMap result suggests that subtypes differ in of small probe molecules binding. This supports the previous finding of having moderate sequence variation across the subtypes for Vpu protein. Hence, it can be assumed that in terms of drug designing subtype specific approach may be needed.

Vpr- Similar process has been followed for Vpr protein for 10 HIV-1 subtypes. According to the result, it is clear that the pockets are more conserved in case of Vpr than both Vif and Vpu. Previously, it has been seen that one particular pocket consisting of specific amino acids may get different ranking in case of Vif and Vpu proteins. That is a difference can be seen for the highest scoring pocket for each subtype predicted by any web server. Whereas in case of the Vpr protein, a similarity has been seen for the highest-ranking pocket along with the constituent amino acids across subtypes predicted by each webserver (**Table 10**). Hence, it can be

concluded that Vpr protein has a better conservancy level across the subtypes. Although, when it comes to a probe binding study, a basic difference can be seen (**Figure 13)**.

Table 10: The pocket residues of the Vpr protein from 10 subtypes. It is quite obvious that the pocket region lies on the helix region which previously proven to be more stable than the terminal regions. The volume of the pocket is also comparable across the subtypes. Though in case of small probe binding, it has been seen that the subtypes can be divided into two distinct groups of regions of highest probe binding cluster.

Subtypes	Predictive pocket Residues from Fpocket, (1st pocket only)	FtMap result of interactive Residues (1st cluster)
VPRA	19,20,22,23,26,27,29,30,39,40,42,43,44,46,47,52,54,56, 57,58,60,61,62,64,65,66,68,69,72	33,34,35,36,68
VPRB	19,20,22,23,26,27,29,30,38,39,40,42,43,44,45,46,47,49,50,52,53,54,56, 57,58,59,60,61,62,63,64,65,66,67,68,69,72	23,27,39,43,52,57,60
VPRC	23,26,27,30,34,38,39,40,42,43,44,45,46,47,49,50,52,53, 56,57,59,60,61,63,64,67	23,27,39,43,52,57,60
VPRD	23,26,27,30,31,37,38,39,40,42,43,44,45,46,47,49,50,52,53,56,57,59,60, 61,63,64,67	33,34,35,36,68
VPRF1	23,26,27,30,31,37,38,39,40,42,43,44,45,46,47,49,50,52,53,56,57,59,60, 61,63,64,67	33,34,35,36,68
VPRF2	18,19,20,22,23,26,27,29,30,38,39,40,42,43,44,45,46,47,49,50,52,53,54, 56,57,58,59,60,61,62,63,64,65,66,67,68,69,72	33,34,35,36,68
VPRG	23,26,27,30,38,39,40,42,43,44,45,46,47,49,50,52,53, 56,57,59,60,61,63,64,67	23,27,39,40,43,52,57,60
VPRH	23,26,27,30,39,40,42,43,44,46,52,53,56,57,59,60,61,63,65,66,68,69,72	23,27,39,43,52,57,60
VPRJ	23,26,27,30,31,37,38,39,40,42,43,44,45,46,47,49,50,52,53,56,57,59,60, 61,62,63,64,65,66,67,68,69,72	23,27,39,43,52,57,60,71
VPRK	19,20,22,23,26,27,29,30,38,39,40,42,43,44,45,46,47,49,50,52,54,56,57, 58,59,60,61,62,63,64,65,66,67,68,69,72	33,34,35,36,68

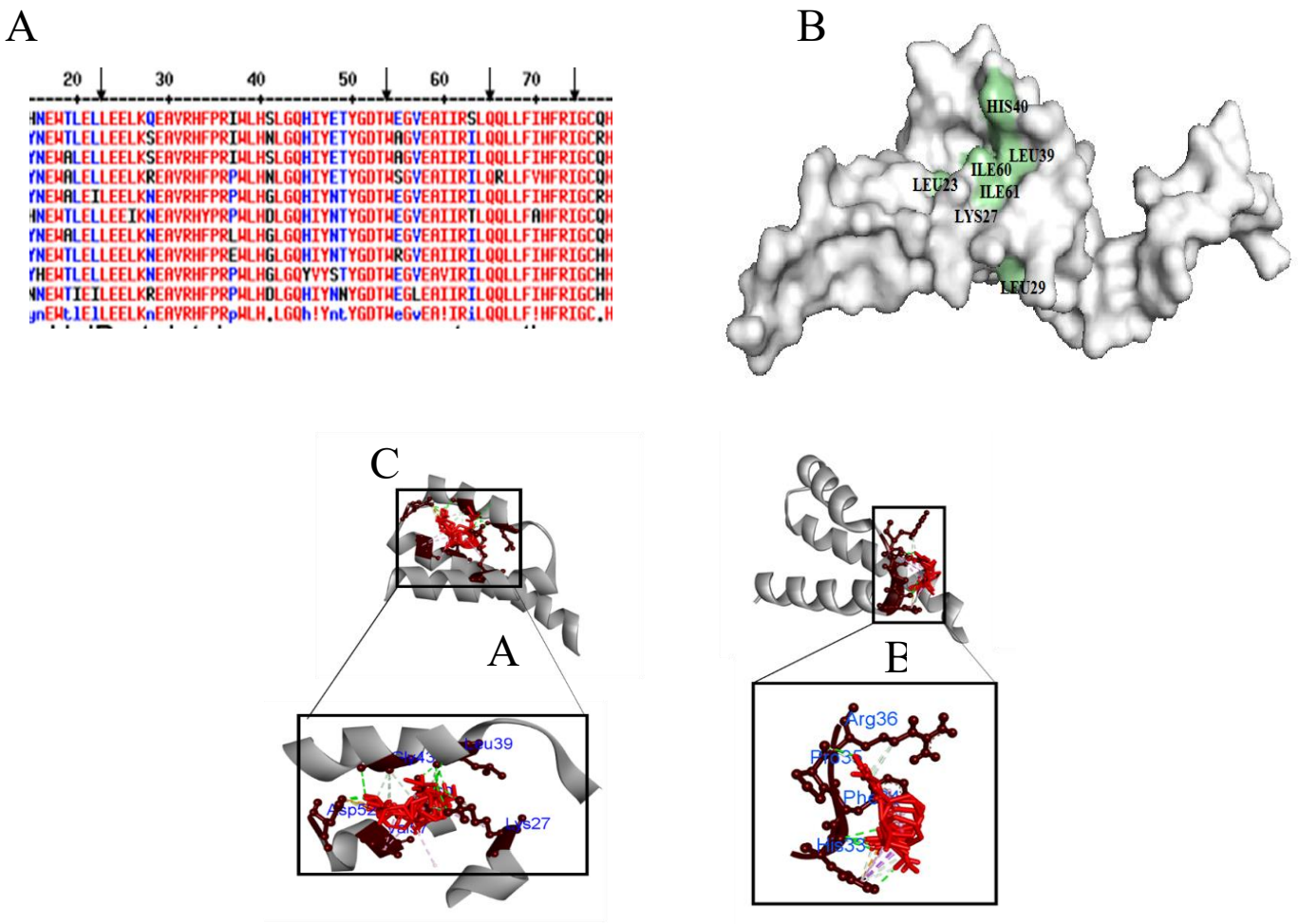
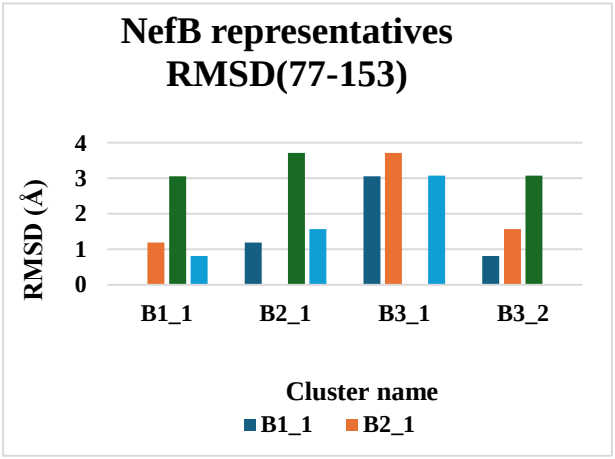


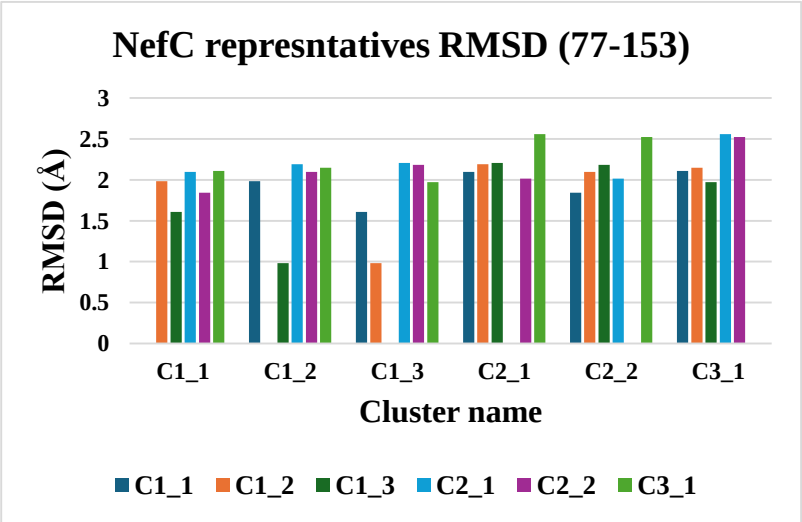
Figure 13: (A) Common pocket region sequences in MSA of all the subtypes. (B) Common pocket of Vpr protein. (C) Two sets of highest probe binding cluster (mentioned Table 7). The main pocket forming amino acid residues are almost identical across the subtypes. According to the FtMap result, the 1st probe binding cluster of 10 subtypes can be divided in to two distinct groups. Moreover, the study has been carried to some more extent to get more details about that pocket and the binding differences among the subtypes. Fragment based designing of small molecules with respect to the specified pocket has been carried out for 10 subtypes of the Vpr protein. A total of 18 different molecules (only the highest energetic molecules are considered) were generated by the AutoGrow 4.0 software for 10 subtypes although the similar parameters

were set for all the subtypes. One question here arises, if the pockets are conserved in terms of both sequence and structure (refer to 1st objective) why a difference is occurring in case of probe binding capacity or fragment based small molecule generation. To explain this a much deeper information needs to be discussed. It is quite clear that the pocket forming residues are conserved at their position but differences can be seen in the adjacent amino acids. This would put versatile effect among the subtypes in case of small molecule binding while keeping the structure same to execute the function. So, it can be concluded that though the subtypes express similar structure to execute the desirable function by protein-protein interaction but they exhibit differences in case of small molecule such as drug like molecule binding.

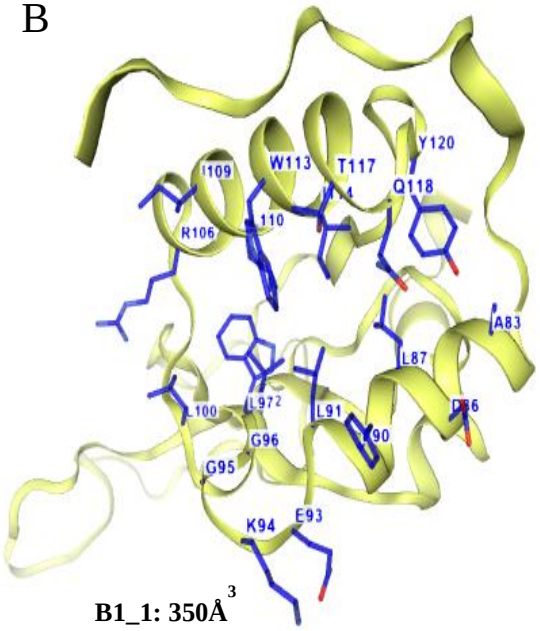
Nef- Referring to the objective 1 result, it can be said that only the region 77-153 retain a stable structure for all the subtypes (**Figure 14**). We have already found that the though blind searching for druggable pocket of Nef protein (pdb :4u5w) predicts 3 prominent pockets but the most appropriate pocket lies between the region 77-153 (preferable 83-120) [92]. The dynamical nature of a particular pocket has been examined by studying the properties (RMSD) of multiple clusters representatives of Nef protein of a particular subtype and across the subtypes. Here, it can be seen that the pocket RMSD differs in a range of 2-3 Å between the cluster of a particular subtype. For inter-subtype this range is even more higher reaches up to a maximum of 8-9Å. Even if the particular pocket differs in size, volume from one cluster representatives to the other of same subtype. Hence, to explore in a more stringent way, a set of four FDA approved drug molecules from ZINC 20 database has been chosen to examine the effect small molecule binding pattern differs across the subtypes in that specific pocket (discussed in the next objective).

A

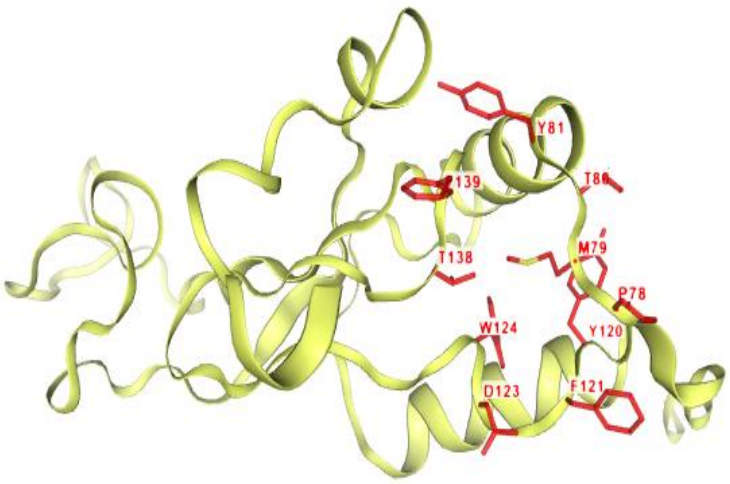




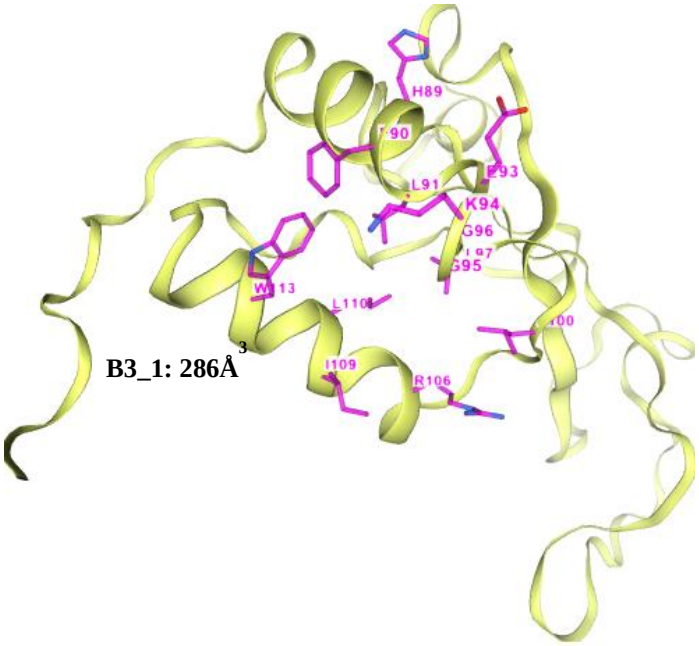
B



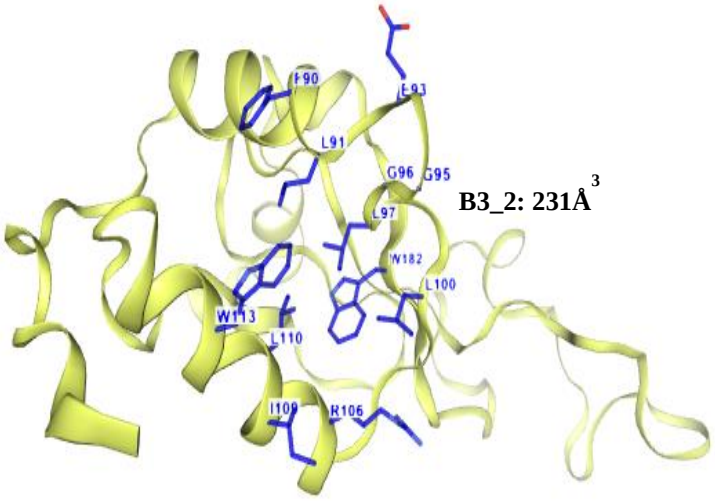
B1_1: 350Å^3



B2_1:191Å^3



B3_1: 286Å^3



B3_2: 231Å^3

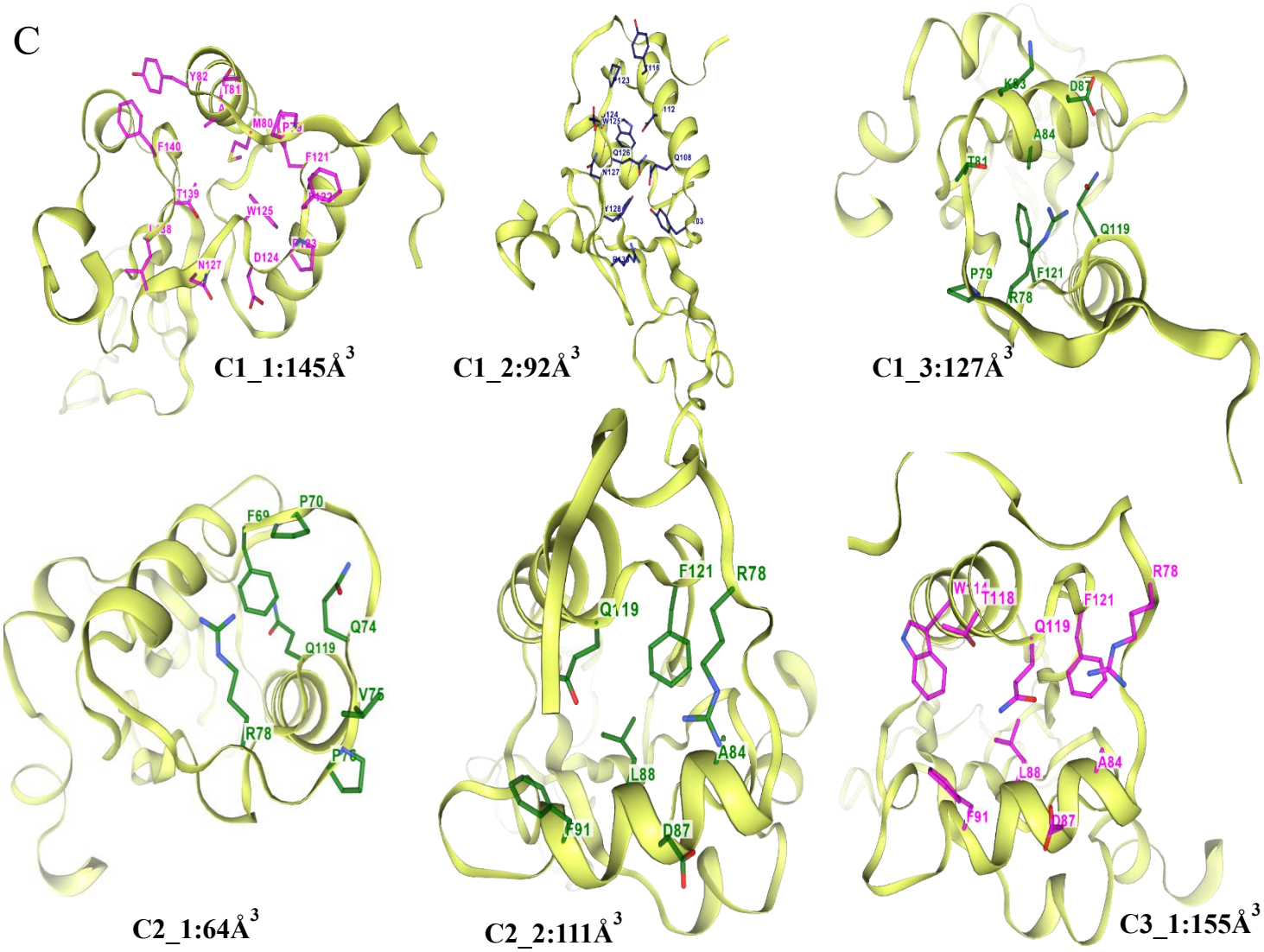


Figure 14: (A) RMSD of region 77-153 of subtype representatives (only subtypes B and C are shown here. Similar studies have been executed for all the subtypes). **(**B and C**)** Stable pockets of subtype B and C representatives respectively. It can be seen that the representatives of subtype B are more conserved having the pocket in the region (83-120) specifically and have comparable pocket volume. Whereas, subtype C representatives have variation in amino acids consisting the specified pocket. For some conformations the pocket consists of slightly different amino acids than 83-120 region. However, the RMSD of the region varies a lot for both the subtypes. Hence, it can be said that the pocket varies from subtypes to subtypes in the aspect of dynamics and geometry as well as amino acids.

OBJECTIVE 3

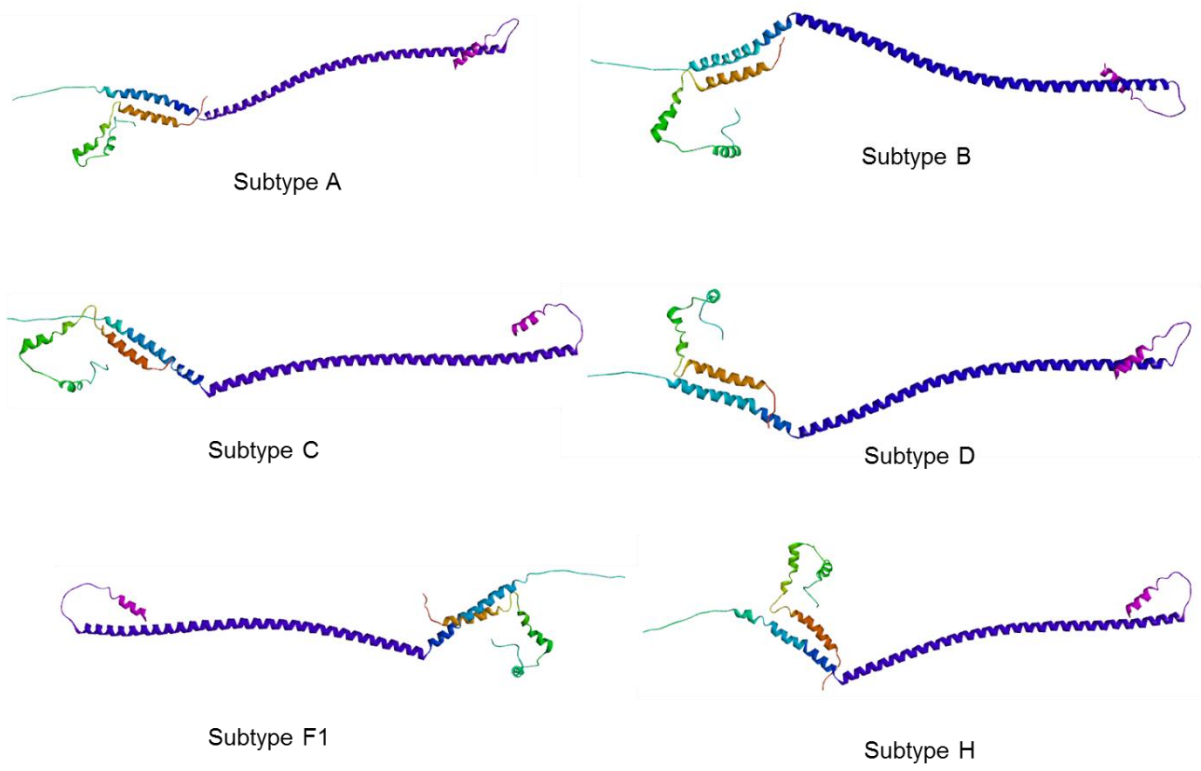
This objective specifically deals with the protein-protein interaction across the subtypes and hotspot area identification. Regarding the result discussed in the 1st objective, it has been seen that the accessory proteins are bound to maintain their structure between the subtypes. Also, in most cases, the amino acids that are proved to be essential for cellular functions of the accessory proteins remain conserved across the subtypes. Here, a more detailed examination of the protein-protein interaction interface of the accessory proteins with their cellular partners has been executed.

Vif- HIV-1 Vif protein interacts with three variants of APOBEC3: APOBEC3G APOBEC3F and APOBEC3H to impede the antiretroviral activity by degradation of the proteins. The interaction profiles of Vif and APOBEC3G, APOBEC3F and APOBEC3H are studied here. It can be said that, A3G interaction domain remain the most conserved region then the A3F region whereas a lot difference has been seen in case of A3H interaction. Thus, subtypic difference is more in the case of A3H interaction than both A3G and A3F (**Table 11**).

Table 11: Various studies have reported the amino acids at the interface of the Vif-A3G, Vif-A3F and Vif-A3H. Most of these studies reported till date were conducted *in vitro* with the constructed Vif gene. Here, the complexes formed by Vif protein of 10 HIV-1 subtypes and different A3 proteins have been studied to determine whether all these subtypes possess the similar interaction pattern with the A3 proteins. 2 servers were used for the docking procedure: ClusPro and Z-Dock. InterProSurf was used to identify the interfaces amino acids between the complexes [93]. From the table, it can be said that the interaction between Vif-A3F is the most conserved across the subtypes. As 10 models for VifB and VifK A3G complex lack the identification from 1st server, hence the confirmation becomes slightly difficult. The least difficult decision to make is for the Vif-A3H interaction across the subtypes. This can be also supported by the fact that a less conservancy has been accounted in the region 60-63 which is the A3H binding region. Both 40-44 region for A3G and 14-17 & 74-79 regions for A3F have higher conservancy level than the aforementioned case.

Subtypes	APOBEC3G (40-44)	APOBEC3H 39,48,60-63	APOBEC3F 11-17,74-79
Vif A	Cluspro, Z-Dock	Cluspro and Z-Dock	Cluspro
Vif B	Z-Dock	Partial identification by Cluspro and Z-Dock	Cluspro and Z-Dock
Vif C	Cluspro, Z-Dock	Cluspro	Cluspro
Vif D	Cluspro, Z-Dock	Cluspro and Z dock	Cluspro and Z-Dock
Vif F1	Cluspro, Z-Dock	Cluspro and Z-Dock	Cluspro and Z-Dock
Vif F2	Cluspro, Z-Dock	Not detected by any of the server	Cluspro and Z-Dock
Vif G	Cluspro, Z-Dock	Not detected by any of the server	Cluspro and Z-Dock
Vif H	Cluspro, Z-Dock	Not detected by any of the server	Cluspro and Z-Dock
Vif J	Cluspro, Z-Dock	Not detected by any of the server	Cluspro and Z-Dock
Vif K	Z-Dock	Cluspro and Z-Dock	Cluspro and Z-Dock

Vpu- Report proves the anti-parallel interaction of the transmembrane domain between the BST2 and HIV-1 Vpu. To investigate the interaction for each subtype, site-specific docking of TMD domain of BST2 and Vpu protein has been executed. Vpu proteins from 6 subtypes has been seen to satisfy the similar interaction involving the important amino acids. However, less dynamic stability has been seen for the protein-protein complexes validated by the MD simulation (**Figure 15**).



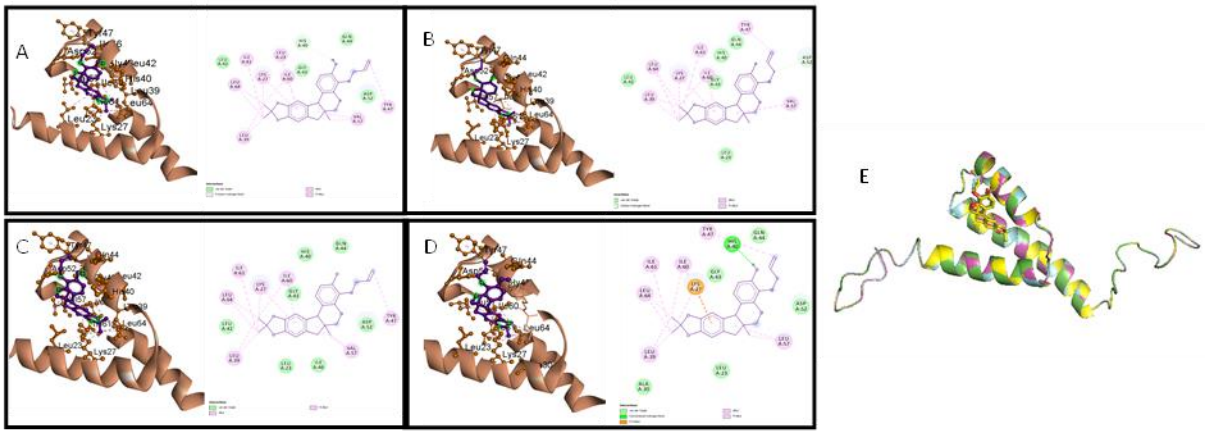
Vpu-BST2 ANTIPARALLEL binding variability

Figure 15: The HADDOCK 2.4 provides several clusters of complexes with different Z score. L11, A18, W22 are the essential amino acids of Vpu protein for this interaction has been identified by several studies. A triad of A10, A14 and A18 (AAA) is also identified as important region for the interaction. According to the MSA result in the 1st objective we have already seen that L11 and W22 remain conserved across the subtypes whereas point mutation has been seen for subtype C (A18V) and F1 (A18T). However, the complexes are not shown to be stable in MD simulation run. Vpu protein got detached from the binding region after 15ns for all the subtypes.

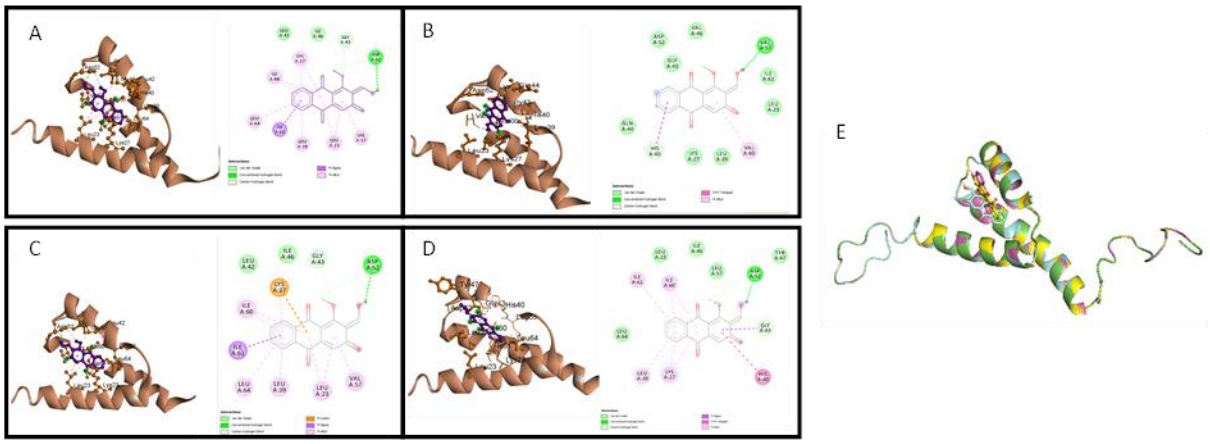
Vpr- Previously it has been determined that mutation in the adjacent position of the pocket forming residues may change the probe or small molecule binding capacity of Vpr protein. Different small molecules have been seen to be generated against the similar binding pocket for each subtype by FBDD. Binding pattern of two sets of molecules: Two naturally occurred Vpr inhibitor (**Figure 16**) and 2 *De Novo* inhibitor like molecules (**Table 12 & Figure 17**) have been studied to determine the differences across the subtypes. The noticeable fact is that, the amino acids involved in the interactions are almost similar for both sets of molecules. Hence,

the binding pocket (amino acids 23,27,39,40,52,57,60,61,62,64,67: the hotspot of PPI interaction of DCAF1-UNG2) is identified as the common Vpr protein drug-applying region for all the HIV-1 subtypes [94].

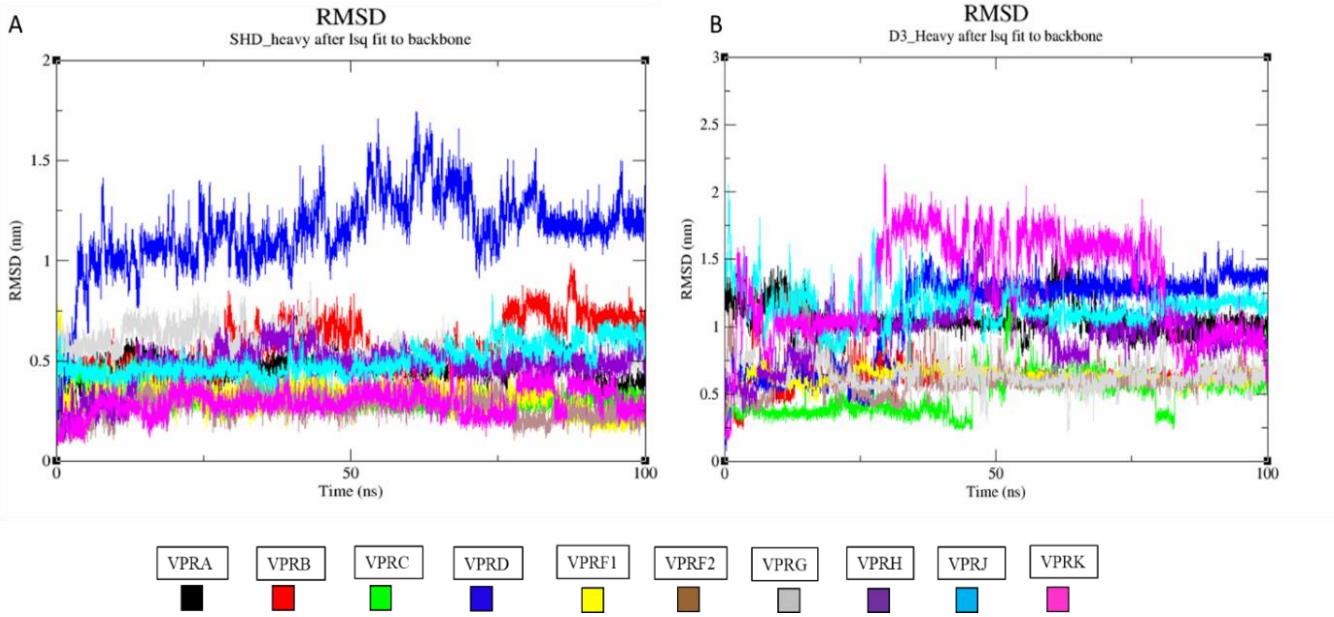
A



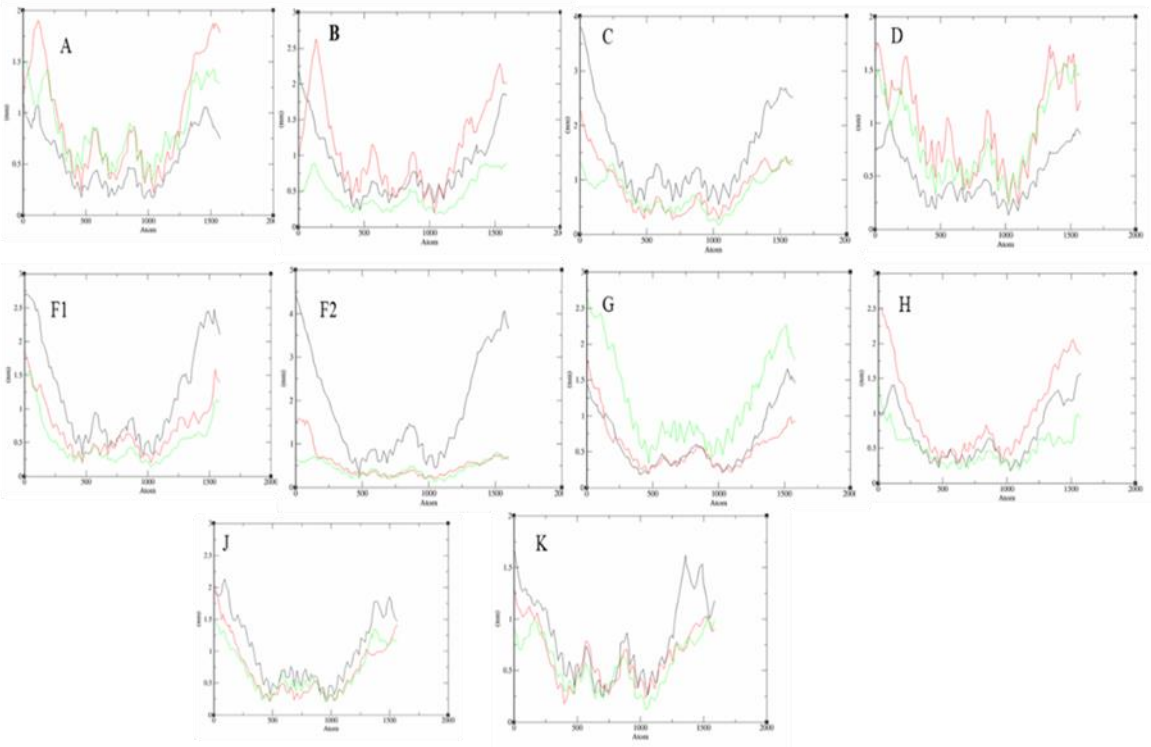
B



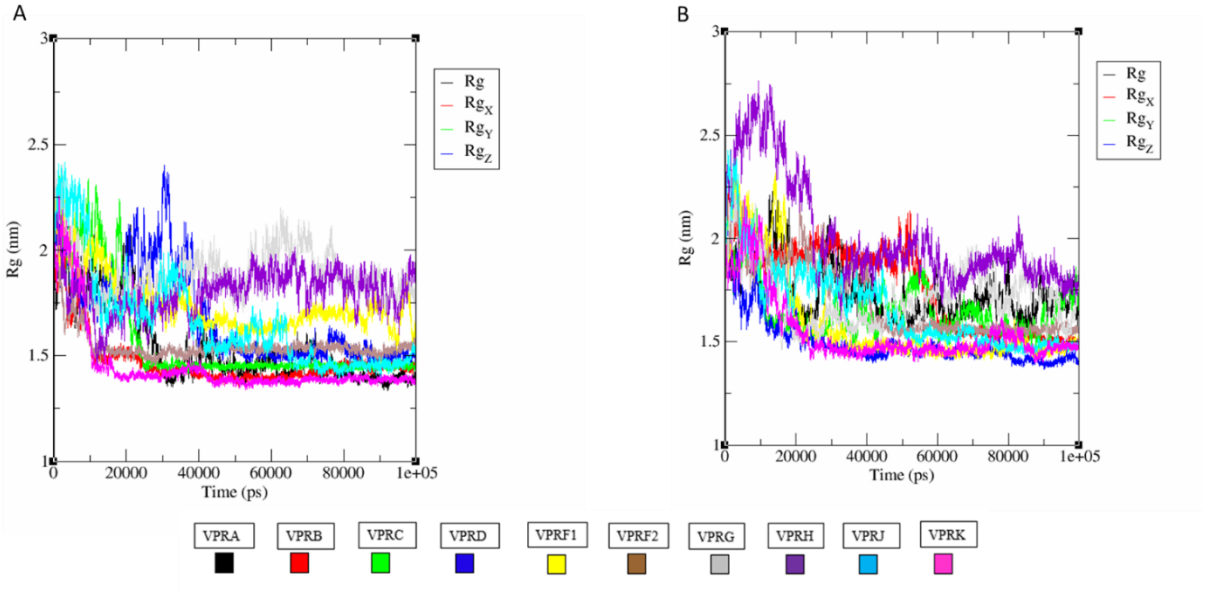
C


A
RMSD
SHD_heavy after lsq fit to backbone
RMSD (nm)
Time (ns)
B
RMSD
D3_Heavy after lsq fit to backbone
RMSD (nm)
Time (ns)
VPRA
VPRB
VPRC
VPRD
VPRF1
VPRF2
VPRG
VPRH
VPRJ
VPRK

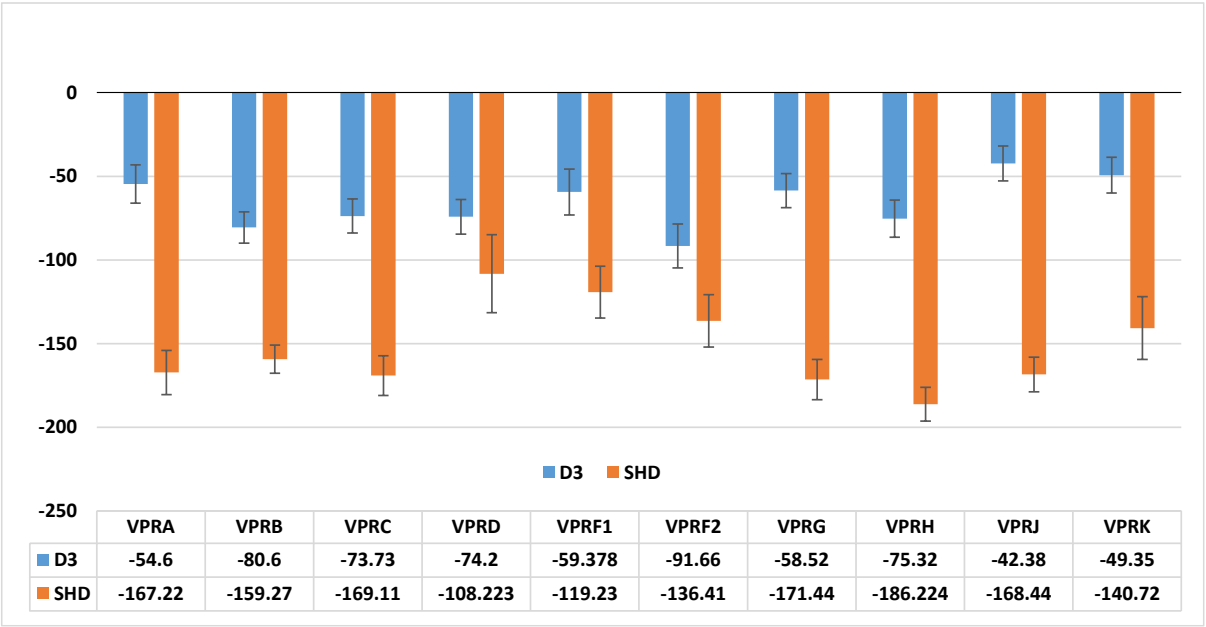
D


A
B
C
D
F1
F2
G
H
J
K

E



F



	VPRA	VPRB	VPRC	VPRD	VPRF1	VPRF2	VPRG	VPRH	VPRJ	VPRK
D3	-54.6	-80.6	-73.73	-74.2	-59.378	-91.66	-58.52	-75.32	-42.38	-49.35
SHD	-167.22	-159.27	-169.11	-108.223	-119.23	-136.41	-171.44	-186.224	-168.44	-140.72

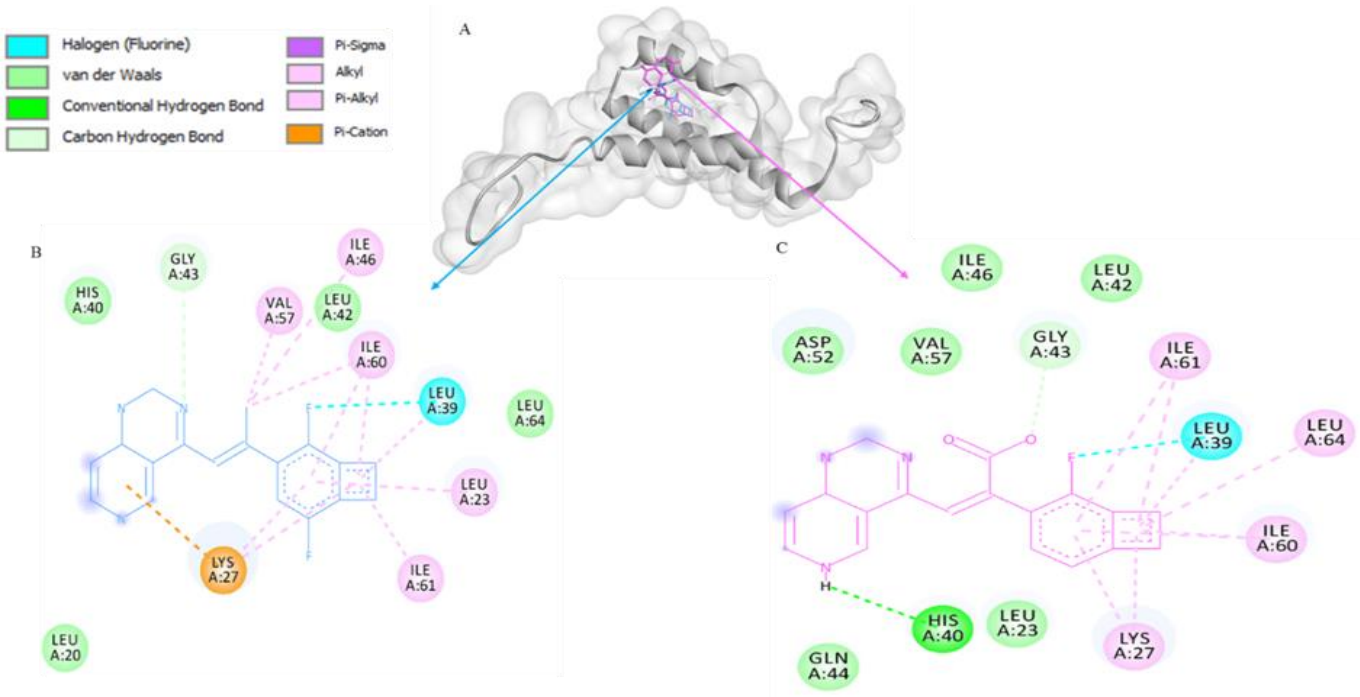
Figure 16A: (A-D) four different poses of SHD (E) SHD binds with Vpr (all subtypes). **16B:** (A-D) four different poses of Damnacanthal (E) Damnacanthal binds with Vpr (all subtypes). **16C:** (A) and (B) Ligand RMSD of Vpr-SHD and Vpr-Damnacanthal protein complexes respectively. **16D:** RMSF of Vpr-SHD (green) and Vpr-Damnacanthal(red). **16E:** R_g of the complexes. **16F:** MM-PBSA energy of the protein-ligand complexes. Damnacanthal and stable hematoxyline derivative (SHD), these 2 naturally occurring Vpr inhibitors are shown to make stable complexes with the Vpr proteins of 10 HIV-1 subtypes. Vpr-SHD complexes reach the plateau after 75ns of trajectory with RMSD 0.4-0.6nm except for the subtype D. Damnacanthal complexes also reach the plateau after 75ns but with a slightly higher ligand RMSD of 1.00nm. The RMSF and R_g graphs also suggest that the complexes are quite stable for Vpr protein of

all subtypes. The molecular contact between the molecules and protein models engages the pocket residues which are also essential for PPI between host and virus. SHD is shown to have stronger binding affinity with Vpr protein than Damnacanthal. Hence, disrupting the PPI (hotspot engaged in protein-inhibitor complexes), these two molecules are proposed to be potential Vpr inhibitors applicable for all the HIV-1 subtypes.

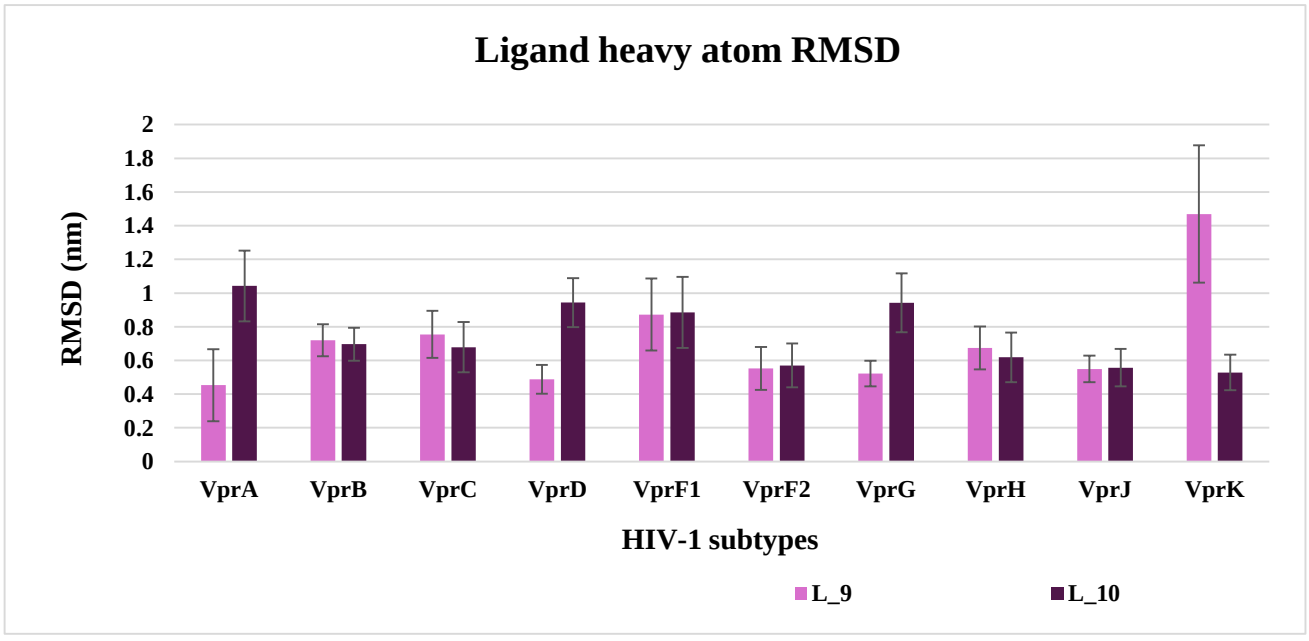
Table 12: AutoDock VINA energy of all the AutoGrow generated molecules from different subtype against common binding pocket. Though, the subtypes generate different molecules themselves but the molecules have comparable binding energy with Vpr protein from all the subtypes. The variation in amino acids adjacent to the pocket forming residues may be the reason for different molecules formation.

Ligands	Origin subtype	HIV-1 Subtypes										
		VprA	VprB	VprC	VprD	VprF1	VprF2	VprG	VprH	VprJ	VprK	Average
L_1	A	-8.8	-8.8	-9.1	-8.8	-8.8	-8.8	-8.9	-8.6	-8.8	-8.9	-8.83
L_2	B	-8.8	-8.9	-9	-8.8	-9.1	-9	-8.9	-9	-9	-8.8	-8.93
L_(3/4)	C	-8.2	-8	-8	-8	-8.1	-8.1	-8	-8	-7.5	-7.5	-7.94
L_5		-8.8	-8.8	-8.8	-8.8	-8.9	-8.8	-8.8	-8.8	-8.1	-7.8	-8.64
L_6	D	-8.3	-8.2	-8.	-8.3	-8.2	-8.3	-8.2	-8.3	-7.6	-8.1	-8.17
L_7		-8.2	-8	-8.1	-8.2	-8.2	-8.2	-8.1	-8.1	-7.4	-7.6	-8.01
L_8		-8.2	-8.2	-8.1	-8.2	-8.2	-8.2	-8	-8.2	-8	-7.2	-8.07
L_9		-8.2	-8	-8.1	-8.2	-8.3	-8.2	-8.2	-8	-7.5	-7.8	-8.04
L_10		-8	-8.2	-8.1	-8.1	-8.2	-8.2	-8.2	-8	-7.2	-7.3	-7.95
L_11		-8.2	-8.2	-8.1	-8.2	-8.2	-8.2	-8.2	-8.2	-8	-7	-8.05
L_12		-8.1	-7.9	-7.9	-8.2	-8.2	-8.2	-8	-7.9	-7.5	-7.3	-7.92
L_(13/14)	F1	-8.3	-8.3	-8.4	-8.3	-8.4	-8.4	8.3	8.4	-8	-7.4	-8.22
L_15		-8.5	-8.4	-8.5	-8.5	-8.5	-8.5	-8.5	-8.5	-8	-7.7	-8.36
L_16	F2	-7.6	-7.6	-7.6	-7.4	-7.5	-7.5	-7.5	-7.5	-8.3	-7.1	-7.56
L_17	G	-7.9	-8	-7.9	-7.9	-8	-8	-8	-8	-8	-7.9	-7.96
L_18	H	-9.2	-8.7	-9.2	-9.1	-9.2	-9.2	-9.2	-9.2	-8.3	-7.5	-8.88
L_19	J	-7.4	-7.5	-7.5	-7.5	-7.5	-7.5	-7.5	-7.5	-8.1	-6.9	-7.49
L_20		-8.3	-8.2	-8.3	-8.3	-8.3	-8.4	-8.3	-8.3	-8.2	-7.2	8.18
L_(21/22)	K	-9.3	-9.3	-9.3	-9.5	-9.4	-9.3	-9.3	-9.4	-8.7	-9.5	-9.32

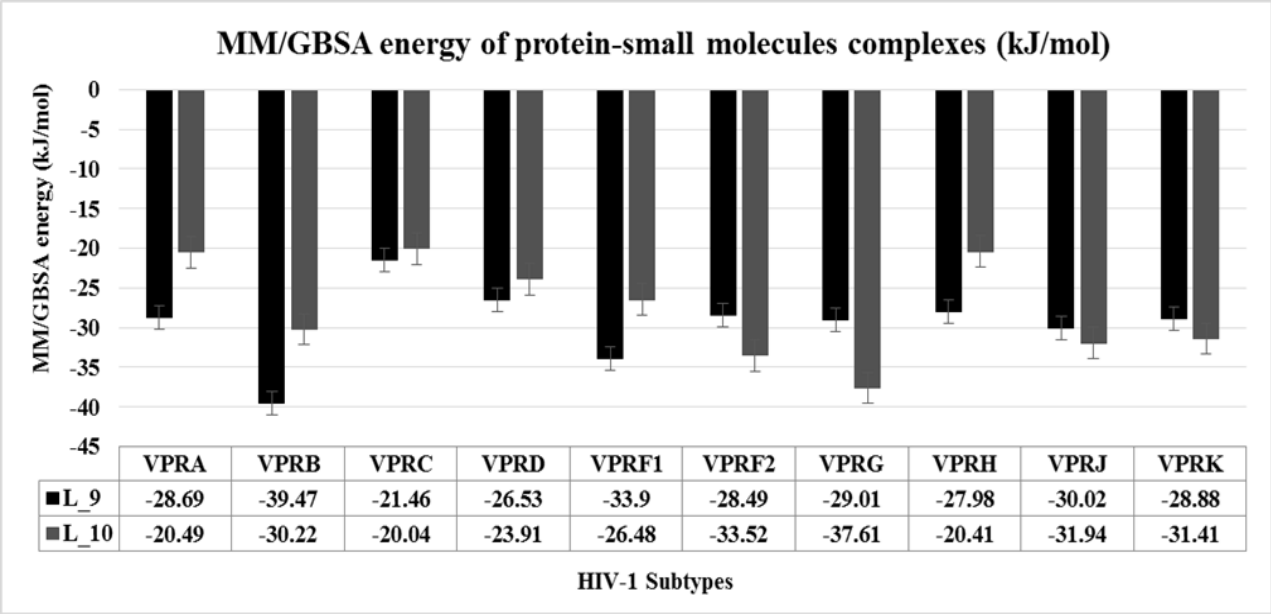
A


Halogen (Fluorine)
van der Waals
Conventional Hydrogen Bond
Carbon Hydrogen Bond
Pi-Sigma
Alkyl
Pi-Alkyl
Pi-Cation
A
B
C
HIS A:40
GLY A:43
ILE A:46
VAL A:57
LEU A:42
ILE A:60
LEU A:39
LEU A:64
LEU A:23
LYS A:27
ILE A:61
LEU A:20
ASP A:52
VAL A:57
GLY A:43
ILE A:46
LEU A:42
ILE A:61
LEU A:39
LEU A:64
ILE A:60
HIS A:40
LEU A:23
LYS A:27
GLN A:44

B


Ligand heavy atom RMSD
RMSD (nm)
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
VprA
VprB
VprC
VprD
VprF1
VprF2
VprG
VprH
VprJ
VprK
HIV-1 subtypes
L_9
L_10

C



	VPRA	VPRB	VPRC	VPRD	VPRF1	VPRF2	VPRG	VPRH	VPRJ	VPRK
L_9	-28.69	-39.47	-21.46	-26.53	-33.9	-28.49	-29.01	-27.98	-30.02	-28.88
L_10	-20.49	-30.22	-20.04	-23.91	-26.48	-33.52	-37.61	-20.41	-31.94	-31.41

Figure 17: (A) Common binding pose of L_9 and L_10 (the orientation differs slightly for subtype J and K, not shown here). (B) RMSD plot of L_9 and L_10 protein complexes. (C) MM/GBSA energy plot of protein-ligand complexes. The CADD generated molecules are also shown to make stable complexes with the Vpr protein of all the subtypes. L_9 RMSD for Subtype K reaches an exceptional ligand RMSD near 1.5nm. For subtypes A, D and G, ligand RMSD of L_9 is almost 0.5 times less than L_10 whereas, for subtypes C, H and K, L_9 has higher RMSD than L_10. The rest of the subtypes have comparable ligand RMSD for both. But, the amino acids involve in the protein-ligand interactions are identical with DCAF1-DDB1-UNG2 binding region. Similarly, MM/GBSA energies of the ligands vary from subtype to subtype. Molecular contact between the molecules and the protein models involves the same pocket forming and hotspot residues. The binding free energies of these amino acids are also determined. The number of H-bond formed in the last 10ns of the simulation trajectory ranges from 3-6.

Nef – Nef-SH3 binding domain is the one of the proved interactions over the years. The most stable pocket consists of this SH3 binding area of Nef protein. Interaction profile of SH3 kinase with the Nef protein from all subtypes have been studied from the protein-protein docked structures. Result shows the similar amino acids and interaction pattern between the subtypes as well as identical Nef protein hot-spot for each subtype (**Table 13**). However, in the binding of small molecules the subtypes behave differently. Subtype B of Nef protein shows a more conservation whereas, subtype C has a scattered pattern. These findings support the previous assumption that the minor sequence variation doesn't affect the protein-protein interaction but shows differences in small molecule binding between the subtypes (**Figure 18**).

Table 13: Hot-spot residues of Nef-SH3 complexes from 10 HIV-1 subtypes by PPCheck webserver [95]. The amino acids are identical for all the cases.

Subtypes	**Res Num**	**Res Name**	**Chain**
NefA	118	TRP	D
	136	TYR	D
	99	ASP	D
NefB	118	TRP	D
	91	ASP	D
	131	TYR	D
NefC	91	ASP	D
NefD	95	ILE	D
	96	HIS	D
	136	TYR	D
	118	TRP	D
NefF1	118	TRP	D
	96	HIS	D
	91	ASP	D
	99	ASP	D
NefF2	95	ILE	D
	118	TRP	D
	96	HIS	D
	136	TYR	D
	92	TYR	D
NefG	95	ILE	D
	96	HIS	D
	136	TYR	D
	118	TRP	D
	133	PRO	D
NefH	95	ILE	D
	118	TRP	D
	96	HIS	D
	136	TYR	D
NefJ	106	ASP	D
Nef K	95	ILE	D
	96	HIS	D
	92	TYR	D
	93	GLU	D
	97	HIS	D

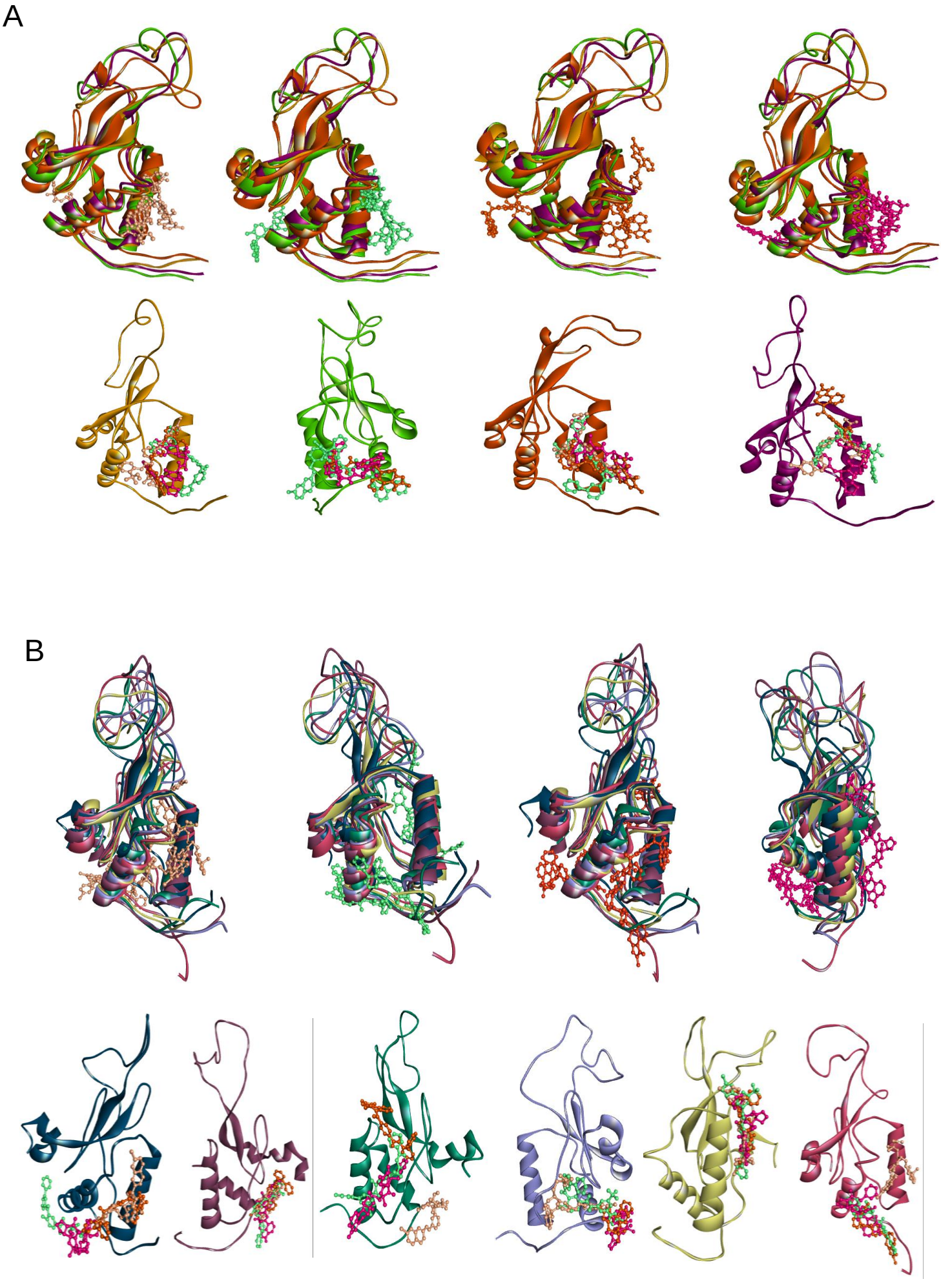

A
B

Figure 18 (A and B): Binding of ZINC 20 database molecules with Nef protein from subtype B (B1_1: brown, B2_1: green, B3_1: crimson and B3_2: purple) and C (C1_1: blue, C1_2: magenta, C1_3: green, C2_1: mauve, C2_2: yellow and C3_1: pink). Four small molecules from ZINC 20 database having the highest binding affinity with all the Nef protein representatives are selected according to the AutoDock Vina energy (Orange: ZINC000003831151, Green: ZINC000006716957, Rust: ZINC000012503187 Pink: ZINC000052955754. In the previous objective, the differences across the subtypes of the pocket have been discussed. Same small molecule has been shown to achieve different binding pose for different subtypes. In case of subtype B, except B2_1, the small molecules bind in the similar binding region (in slightly different conformation) for other three representatives. Whereas, for subtype C the molecules are seen to possess different binding positions for different representatives. On the other hand, for any particular representatives of subtype B, all the molecules bind almost in the identical region. In some cases of Subtype C representatives, the molecules have completely different position than the most probable region. This study has been executed for rest of the subtypes also.

CONCLUSION

In this study, the similarity and dissimilarity between the 10 subtypes of HIV-1 have been examined considering 4 accessory proteins (Nef, vif, vpr and Vpu) of HIV-1 as system. MSA has been executed to the accessory proteins of 10 subtypes individually with the data from LANL database. Starting from the sequence level, it is shown that Nef and Vif proteins respectively have the highest and lowest sequence variation between the subtypes. It has been also seen that the virus conserves the amino acids which are essential for the protein to play its respective action for most of the cases. Even if the amino acid gets altered the mutated amino acid belongs to the similar type. However, 25-30% sequence variability is present across the subtypes for all the proteins. Clustering result with a 75% similarity cut off shows that Nef and Vpu sequences divided in to several clusters whereas, Vif and Vpr sequences belong to one single cluster (either subtype B or C). Shannon entropy data also supports the idea that Vif and Vpr proteins are much more conserved in amino acid position than Nef and Vpu. So, it can be concluded that sequence variation is there between the subtypes of HIV-1 accessory proteins. There is no common trend has been shown in phylogenetic analysis among the subtypes with the protein sequences except some facts that subtype D has diverged from the other clades and has subtype G as the nearest neighbor. Subtype B &C and subtype F1&F2 are diverged from common ancestor with very low distance value. The virus maintains the secondary structures of the accessory proteins across the subtypes but differences are shown in the superimposed tertiary structures. Again, Nef protein differ in highest extent, Vpu protein differs in the terminal regions whereas, tertiary structure of Vif and Vpr proteins are very well superimposed (referencing the RMSD results). However, difference in the lowest energy structures have been determined for Vpr protein across the subtypes. Using this cluster representatives (with different conformations) for each subtype of Nef protein, the dynamics of a particular druggable pocket region has been examined. It has been seen that subtype B has a more conservation in pocket region across the representatives whereas the representatives of subtype C varies a lot more. Even the amino acids consisting the pocket region varies among the representatives. Hence, Nef proteins show differences in pocket dynamics and geometry across the subtypes. However, a much-conserved results for druggable pockets are seen for the other 3 proteins Vif, Vpr and Vpu. These proteins have common pocket regions (consisting of similar amino acids) which sometimes differ in properties like druggability, volume or score from subtype to subtype. Nevertheless, differences have been found across the subtypes in case of small probe binding capacity for all the accessory proteins except Vif protein. For Vpr proteins, subtypes got divided in to two distinct groups of having highest binding cluster at two different

regions. Vpu proteins from different 6 subtypes of HIV-1 have almost different regions of highest probe binding cluster. Hence, it can be assumed that though the druggable pockets of accessory proteins remain conserved across the subtypes they can differ in binding of small molecules. Vif being the most conserved accessory protein it maintains similar interaction pattern to A3G, A3H and A3F across the subtypes. The docked structures of Vpu-BST2 confirmed the similar interaction between the proteins for all the subtypes. Identical interaction hotspots are also identified for Nef–SH3 binding for all the subtypes. Vpr protein has conserved UNG2 domain across the subtypes and the interactive amino acids are similar for most of subtypes. Binding of small molecules in the PPI region (common pocket region also) with Vpr protein show identical molecular contact and comparable post simulation analysis data between the subtypes. However, for Nef protein, similar binding pattern of one particular molecule has been shown for the different conformations of subtype B but have large differences for the conformations of subtypes C. In conclusion it can be said that though the subtypes maintain the similar binding pattern and molecular contact in large molecule interaction like protein-protein interaction they may differ in binding of small molecules. According to the result, it can be suggested that generalized drug approach can be relevant for Vpr protein but subtype-specific approach would have been beneficial targeting Nef protein in drug development research.

SIGNIFICANCE OF THE STUDY

It always remains a question that whether there is requirement for subtypic specific therapeutics for HIV-1 treatment or any generalized therapeutics can work. In this study, the similarity and dissimilarity between the subtypes are evaluated in detail emphasizing on the druggable pocket and PPI region. It is shown that despite of having identical druggable pocket, difference is there in case of small molecule binding. Being the less conserved protein, Nef expresses multiple conformations with high RMSD value. On the other hand, Vpr protein having higher conservancy, less difference has been seen for small molecule binding. Later, it has been seen that having differences between the subtypes Vpr protein shows a common approach applicable to all subtypes. Both the set of molecules shown to share identical binding pose, molecular contact, involvement of individual amino acids of PPi hotspot and almost comparable binding energy across the subtypes. So, subtype generalized approach may be fitted to therapy targeting HIV-1 Vpr protein. On the other side, Nef protein shows multiple conformations with higher RMSD values in any one subtypes as well as among the subtypes. Dynamic changes have been spotted for the druggable pocket across the subtypes. Differences has been spotted out in binding pose and orientation and molecular contact of similar drug like molecules to different subtypes of Nef protein. This study provides a detail insight how the accessory proteins of HIV-1 behave in both protein-protein and protein-small molecule interaction considering similarity and dissimilarity from sequence to structural level. This will help in designing potential inhibitor against the HIV-1 accessory proteins in future.

FUTURE DIRECTION

The repurposing of the ZINC molecules targeting the Nef protein can be validated by studying the protein-ligand complex stability applying MD simulation and post simulation analysis. Examine the dynamic stability of the Vif-A3 complexes can be validated using long-range MD simulation study. This current study failed to establish to stability of the Vpu-BST2 complexes. Conducting a long-range simulation run may help in finding the actual binding propensity scenario between these two proteins. These investigations will provide comprehensive idea about the number of bond formation-breaking between the protein partners, conformational changes of the accessory proteins, stability of the protein-protein complexes. Depending on the data, it will be easier to suggest the appropriate strategy of applying either subtype-specific or generalized approach in drug development targeting the interruption in PPI region.

REFERENCES

1. Jia XC, Xia ZH, Shi N, et al (2020). The factors associated with natural disease progression from HIV to AIDS in the absence of ART, a propensity score matching analysis. *Epidemiol Infect*.; 148:e57. https://doi:10.1017/S0950268820000540

2. Shaw GM, Hunter E. (2012). HIV transmission. *Cold Spring Harb Perspect Med*., 2 (11):a006965. https://doi:10.1101/cshperspect.a006965.

3. Rawson JM, Landman SR, Reilly CS, Mansky LM (2015). HIV-1 and HIV-2 exhibit similar mutation frequencies and spectra in the absence of G-to-A hypermutation. *Retrovirology*, 12(60). https:doi: 10.1186/s12977-015-0180-6

4. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM (2008). The challenge of HIV-1 subtype diversity. *N Engl J Med.*, 358(15):1590-1602. https://doi: 10.1056/NEJMra0706737

5. Li G, Piampongsant S, Faria NR, et al. (2015). An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology*, 12:18. https://doi.org/10.1186/s12977-015-0148-6

6. Olabode AS, Ng GT, Wade KE, et al (2022). Revisiting the recombinant history of HIV-1 group M with dynamic network community detection. *Proc Natl Acad Sci U S A*, 119:e2108815119. https://doi.org/10.1073/pnas.2108815119

7. Ndjoyi-Mbiguino A, Zoa-Assoumou S, Mourembou G, Ennaji MM (2020). Chapter 10 - Human Immunodeficiency Virus: A Brief Review. *Emerging and Reemerging Viral Pathogens*, pp 183-200. https://doi.org/10.1016/B978-0-12-819400-3.00010-7

8. Golmohammadi R, Baesi K, Moradi A, Farrokhi M, McFarland W, Parsamajd S (2017). The First Characterization of HIV-1 Subtypes and Drug Resistance Mutations among Antiretrovirally Treated Patients in Kermanshah, Iran. *Intervirology*. , 60(1-2):33-37. https://doi: 10.1159/000478701

9. Chaplin B, Eisen G, Idoko J, Onwujekwe D, Idigbe E, Adewole I, Gashau W, Meloni S, Sarr AD, Sankalé JL, Ekong E, Murphy RL, Kanki P (2011). Impact of HIV type 1 subtype on drug resistance mutations in Nigerian patients failing first-line therapy. *AIDS Res Hum Retroviruses*, 27(1):71-80. https://doi: 10.1089/aid.2010.0050

10. Kantor R, Katzenstein DA, Efron B, Carvalho AP, Wynhoven B, et al (2005). Impact of HIV-1 Subtype and Antiretroviral Therapy on Protease and Reverse Transcriptase Genotype:

Results of a Global Collaboration. *PLOS Medicine,* 2(4):e112. https://doi.org/10.1371/journal.pmed.0020112

11. Wensing AM, Calvez V, Ceccherini-Silberstein F, Charpentier C, Günthard HF, Paredes R, Shafer RW, Richman DD (2019). 2019 update of the drug resistance mutations in HIV-1. *Top Antivir Med.*, 27(3):111-121.

12. Lu X, Zhao H, Zhang Y *et al* (2017). HIV-1 drug-resistant mutations and related risk factors among HIV-1-positive individuals experiencing treatment failure in Hebei Province, China. *AIDS Res Ther*., 14:4. https://doi.org/10.1186/s12981-017-0133-3

13. German Advisory Committee Blood (Arbeitskreis Blut), Subgroup 'Assessment of Pathogens Transmissible by Blood' (2016) Human Immunodeficiency Virus (HIV). *Transfus Med Hemother,* 43:203-222. https://doi.org/10.1159/000445852

14. Addo MM , Altfeld M, Rosenberg ES, Eldridge RL, Philips MN, Habeeb K, Khatri A, Brander C, Robbins GK, Mazzara GP, Goulde r PJR, Walker BD, and the HIV Controller Study Collaboration (2001). The HIV-1 regulatory proteins Tat and Rev are frequently targeted by cytotoxic T lymphocytes derived from HIV-1-infected individuals. *PNAS*, 98(4):1781-1786. https://doi.org/10.1073/pnas.98.4.1781

15. Engelman A, Cherepanov P (2012). The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol.*, 10:279–290. https://doi.org/10.1038/nrmicro2747

16. Strebel K. (2013) HIV accessory proteins versus host restriction factors. *Current opinion in virology*, 3:692–699. https://doi.org/10.1016/j.coviro.2013.08.004

17. Zotova A, Atemasova A, Pichugin A, Filatov A, Mazurov D (2019). Distinct Requirements for HIV-1 Accessory Proteins during Cell Coculture and Cell-Free Infection. *Viruses*, 11(5):390. https://doi:10.3390/v11050390

18. Azimi FC, Lee JE (2020). Structural perspectives on HIV-1 Vif and APOBEC3 restriction factor interactions. *Protein Sci*. 29(2):391-406. doi:10.1002/pro.3729

19. Binka M, Ooms M, Steward M, Simon V. The activity spectrum of Vif from multiple HIV-1 subtypes against APOBEC3G, APOBEC3F, and APOBEC3H. *J Virol*. 2012;86(1):49-59. https://doi:10.1128/JVI.06082-11

20. Zhao RY, Bukrinsky MI (2014). HIV-1 accessory proteins: VpR. *Methods Mol Biol.* ,1087:125-134. https://10.1007/978-1-62703-670-2_11

21. González ME (2017). The HIV-1 Vpr Protein: A Multifaceted Target for Therapeutic Intervention. *Int J Mol Sci.,* 18:126. https://doi.org/10.3390/ijms18010126

22. Miyatake H, Sanjoh A, Murakami T, Murakami H, Matsuda G, Hagiwara K, Yokoyama M, Sato H, Miyamoto Y, Dohmae N, Aida Y (2016). Molecular Mechanism of HIV-1 Vpr for

Binding to Importin-α, *J.Mol.bio*., 428(2016):2744-2757. https://doi.org/10.1016/j.jmb.2016.05.003

23. Wu Y, Zhou X, Barnes CO, et al. (2016). The DDB1-DCAF1-Vpr-UNG2 crystal structure reveals how HIV-1 Vpr steers human UNG2 toward destruction. *Nat Struct Mol Biol*. ,23:933-940. https://doi.org/10.1038/nsmb.3284

24. Sato H, Murakami T, Matsuura R, et al. (2022). A Novel Class of HIV-1 Inhibitors Targeting the Vpr-Induced G2-Arrest in Macrophages by New Yeast- and Cell-Based High-Throughput Screening. *Viruses*, 14:1321. https://doi.org/10.3390/v14061321

25. Kogan M, Rappaport J (2011). HIV-1 accessory protein Vpr: relevance in the pathogenesis of HIV and potential for therapeutic intervention. *Retrovirology*, 8:25. https://doi.org/10.1186/1742-4690-8-25

26. Jia X, Weber E, Tokarev A, et al. (2014). Structural basis of HIV-1 Vpu-mediated BST2 antagonism via hijacking of the clathrin adaptor protein complex 1. *Elife*, 3:e02362. https://doi:10.7554/eLife.02362

27. Manninen A, Hiipakka M, Vihinen M, Lu W, Mayer BJ, Saksela K (1988). SH3-Domain binding function of HIV-1 Nef is required for association with a PAK-related kinase. *Virology*., 250(2):273-282. https://doi:10.1006/viro.1998.9381

28. Arts EJ, Hazuda DJ (2012). HIV-1 antiretroviral drug therapy. *Cold Spring Harb Perspect Med.,* 2(4):a007161. https://doi:10.1101/cshperspect.a007161

29. Wainberg MA, Brenner BG (2012). The Impact of HIV Genetic Polymorphisms and Subtype Differences on the Occurrence of Resistance to Antiretroviral Drugs. *Mol Biol Int*., 2012:256982. https://doi:10.1155/2012/256982

30. Singh K, Flores JA, Kirby KA, et al. (2014). Drug resistance in non-B subtype HIV-1: impact of HIV-1 reverse transcriptase inhibitors. *Viruses*., 6(9):3535-3562. https://doi:10.3390/v6093535

31. HIV Sequence Compendium 2021 Apetrei C, Hahn B, Rambaut A, Wolinsky S, Brister JR, Keele B, Faser C, Eds. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR-23-22840.

32. UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51:D523-D531. https://doi.org/10.1093/nar/gkac1052

33. Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, Notredame C (2016). Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.*, 17(6):1009–1023, https://doi.org/10.1093/bib/bbv099

34. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191. https://doi.org/10.1093/bioinformatics/btp033

35. Chiu JKH, Ong RTH (2022). Clustering biological sequences with dynamic sequence similarity threshold. *BMC Bioinformatics,* **23**:108. https://doi.org/10.1186/s12859-022-04643-9

36. Sen S, Dey A, Chowdhury S. *et al.* (2019). Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis. *BMC Bioinformatics,* **19**(Suppl13):549. https://doi.org/10.1186/s12859-018-2552-0

37. Buchan DWA, Jones DT (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.*, 47(W1):W402–W407, https://doi.org/10.1093/nar/gkz297

38. Waterhouse A, Bertoni M, Bienert S, et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*., 46:W296-W303. https://doi.org/10.1093/nar/gky427

39. Agnihotry S, Pathak RK, Singh DB, Tiwari A, Hussain I (2022). Chapter 11 - Protein structure prediction, Bioinformatics., pp177-188.

40. An J, Totrov M, Abagyan R (2004). Comprehensive identification of "druggable" protein ligand binding sites. *Genome informatics*. International Conference on Genome Informatics.15:31-41. https://doi.org/10.11234/gi1990.15.2_31

41. Hajduk PJ, Huth JR, Tse C (2005). Predicting protein druggability. *Drug Discov Today*. 10:1675-1682. https://doi.org/10.1016/S1359-6446(05)03624-X

42. Le Guilloux V, Schmidtke P, Tuffery P (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics.* 10:168. https://doi.org/10.1186/1471-2105-10-168

43. Schmidtke P, LE Guilloux V, Maupetit J, Tuffery P (2010). fpocket: online tools for protein ensemble pocket detection and tracking, *Nucleic Acids Res.*, 38(suppl_2):W582–W589. https://doi.org/10.1093/nar/gkq383

44. Kochnev Y, Durrant JD (2022). FPocketWeb: protein pocket hunting in a web browser. *J Cheminform*, 14:58. https://doi.org/10.1186/s13321-022-00637-0

45. Schöning-Stierand K, Diedrich K, Fährrolfes R, et al. (2020) ProteinsPlus: interactive analysis of protein-ligand binding interfaces. *Nucleic Acids Res*., 48:W48-W53. https://doi.org/10.1093/nar/gkaa235

46. Volkamer A, Kuhn D, Rippmann F, Rarey M(2012). DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics*, 28:2074-2075. https://doi.org/10.1093/bioinformatics/bts310

47. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33:3036-3042.https://doi.org/10.1093/bioinformatics/btx350

48. Yang Liu, Maximilian Grimm, et al (2019). CB-Dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacologica Sinica. 41*(1), 138–144. https://doi.org/10.1038/s41401-019-0228-6

49. Yang Cao, Lei Li (2014). Improved protein-ligand binding affinity prediction by using a curvature dependent surface area model. *Bioinformatics, 30*(12):1674–1680. https://doi.org/10.1093/bioinformatics/btu104

50. Yang Liu, et al (2022). CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Res*., *50*(W1):W159–W164. https://doi.org/10.1093/nar/gkac394

51. Kozakov D, Grove LE, Hall DR, et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc,* 10:733-755. https://doi.org/10.1038/nprot.2015.043

52. Brenke R, Kozakov D, Chuang GY, et al. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics,* 25:621-627. https://doi.org/10.1093/bioinformatics/btp036

53. Kozakov D, Hall DR, Chuang GY, et al. (2011). Structural conservation of druggable hot spots in protein-protein interfaces. *Proc Natl Acad Sci U S A.,* 108:13528-13533. https://doi.org/10.1073/pnas.1101835108

54. Bohnuud T, Beglov D, Ngan CH, et al. (2012). Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.,* 40:7644-7652. https://doi.org/10.1093/nar/gks519

55. Spiegel JO, Durrant JD (2020). AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J Cheminform,* 12:25. https://doi.org/10.1186/s13321-020-00429-4

56. Trott O, Olson AJ (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31:455-461. https://doi.org/10.1002/jcc.21334

57. Sterling T, Irwin JJ (2015). ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model*, 55:2324-2337. https://doi.org/10.1021/acs.jcim.5b00559

58. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011). Open Babel: An open chemical toolbox. *J Cheminform,* 3:33. https://doi.org/10.1186/1758-2946-3-33

59. Forli, S., Huey, R., Pique, M. *et al.* (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat Protoc,* 11:905–919. https://doi.org/10.1038/nprot.2016.051

60. Hagiwara K, Ishii H, Murakami T, Takeshima Sn, Chutiwitoonchai N, et al. (2015). Synthesis of a Vpr-Binding Derivative for Use as a Novel HIV-1 Inhibitor. *PLOS ONE*, 10(12):e0145573. https://doi.org/10.1371/journal.pone.0145573

61. Win, N.N., Ngw,e H., Abe, I., Morita, H. (2017). Naturally occurring Vpr inhibitors from medicinal plants of Myanmar. *J Nat Med.*, 71(4):579-589. https://doi: 10.1007/s11418-017-1104-7.

62. Woo S, Win NN, Wyine MNOo, Ngwe H, Ito T, Abe I, Morita H (2019) . Viral protein R inhibitors from Swertia chirata of Myanmar. *Journal of Bioscience and Bioengineering*. ,128(4):445-449. https://doi.org/10.1016/j.jbiosc.2019.04.006

63. Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling. 61*(8):3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

64. Trott O, & Olson A J (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*., 31(2):455-461. https://doi:10.1002/jcc.21334

65. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem,*, *30*(16):2785–2791. https://doi.org/10.1002/jcc.21256

66. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA (2020). ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery, *Journal of Chemical Information and Modeling*, 60 (12):6065-6073. https://DOI: 10.1021/acs.jcim.0c00675

67. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*., 30(12):1771-1773. https://doi:10.1093/bioinformatics/btu097

68. Pierce BG, Hourai Y, Weng Z (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library *PLoS One*., 6(9):e24657. https://doi:10.1371/journal.pone.0024657

69. Kozakov D, Hall DR, Xia B, et al. (2017). The ClusPro web server for protein-protein docking. *Nat Protoc*.,12(2):255-278. https://doi:10.1038/nprot.2016.169

70. Honorato RV, Trellet ME, Jiménez-García1 B, Schaarschmidt JJ, Giulini M, Reys V, Koukos PI, Rodrigues JPGLM, Karaca E, van Zundert GCP, Roel-Touris J, van Noort CW, Jandová Z, Melquiond ASJ, Bonvin AMJJ (2024). The HADDOCK2.4 web server: A leap forward in integrative modelling of biomolecular complexes. *Nature Prot.* https://doi:10.1038/s41596-024-01011-0

71. Honorato RV, Koukos PI, Jimenez-Garcia B, Tsaregorodtsev A, Verlato M, Giachetti A, Rosato A, Bonvin AMJJ (2021). Structural biology in the clouds: The WeNMR-EOSC Ecosystem. *Frontiers Mol. Biosci.,* 8:729513. https://doi:10.3389/fmolb.2021.729513

72. Adcock SA, McCammon JA (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev*., 106:1589-1615. https://doi.org/10.1021/cr040426m

73. Bussi G, Donadio D, Parrinello M (2007). Canonical sampling through velocity rescaling. *J Chem*., 126:014101. https://doi: 10.1063/1.2408420

74. Tuble SC, Anwar J, Gale JD (2004). An approach to developing a force field for molecular simulation of martensitic phase transitions between phases with subtle differences in energy and structure. *J Am Chem Soc*, 126:396-405. https://doi.org/10.1021/ja0356131

75. Darden T, York D, Pedersen L (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.,* 98:10089–10092 https://doi.org/10.1063/1.464397

76. Wu EL, Cheng X, Rui S Jo H, Song KC, Dávila-Contreras EM, Qi Y, Lee J, Monje-Galvan V, Venable RM, Klauda JB, Im W (2014). CHARMM-GUI Membrane Builder Toward Realistic Biological Membrane Simulations. *J. Comput. Chem*. 35:1997-2004. https://doi.org/10.1002/jcc.23702

77. Lindahl, Abraham, Hess, & van der Spoel. (2021). GROMACS 2020.5 Manual (2020.5). Zenodo.

78. Lindahl, Abraham, Hess, & van der Spoel. (2021). GROMACS 2020.5 Source code (2020.5). Zenodo

79. Sousa da Silva AW, Vranken WF (2012). ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* 5:367. https://doi.org/10.1186/1756-0500-5-367

80. Izadi S, Anandakrishnan R, Onufriev AV (2014). Building Water Models: A Different Approach. *J Phys Chem Lett* 5:3863-3871. https://doi.org/10.1021/jz501780a

81. Bjelkmar P, Larsson P, Cuendet MA, Hess B, Lindahl E (2010). Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *J. Chem. Theory Comput*., 6 (2):459-466. https://doi: 10.1021/ct900549r

82. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved Side-Chain Torsion Potentials for the Amber Ff99sb Protein Force Field. *Proteins: Structure, Function, and Bioinformatics* 78:1950-1958. http://dx.doi.org/10.1002/prot.22711

83. Cole GB, Sharpe S (2019). Hydrophobic matching of HIV-1 Vpu transmembrane helix-helix interactions is optimized for subcellular location. *Biochimica et Biophysica Acta (BBA) – Biomembranes*.1861(10). https://doi.org/10.1016/j.bbamem.2019.07.010

84. Majumder S, Giri K (2021). An insight into the binding mechanism of Viprinin and its morpholine and piperidine derivatives with HIV-1 Vpr: molecular dynamics simulation, principal component analysis and binding free energy calculation study. *J Biomol Struct Dyn*., 23:1-13. *https://doi 10.1080/07391102.2021.1954553*

85. Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JH, Noé F (2015). PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput*.,11(11):5525-42. https://doi: 10.1021/acs.jctc.5b00743

86. Kumari et al. (2014). g_mmpbsa - A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.,* 54(7):1951-1962. https://doi.org/10.1021/ci500020m

87. Wang C, Nguyen PH, Pham K, Huynh D, Le TB, Wang H, Ren P, Luo R (2016). Calculating protein-ligand binding affinities with MMPBSA: Method and error analysis. *J Comput Chem*., 37(27):2436-2446. *https://doi: 10.1002/jcc.24467*

88. Baker et al (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci*., 98:10037-10041. https://doi.org/10.1073/pnas.181342398

89. Genheden S, Ryde U (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov*, 10:449-461. https://doi.org/10.1517/17460441.2015.1032936

90. Valdés-Tresanco MS, Valdés-Tresanco ME, Valiente PA, Moreno E (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *J Chem Theory Comput* 17:6281-6291. https://doi.org/10.1021/acs.jctc.1c00645

91. Miller BR 3rd, McGee TD Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput 8:3314-3321. https://doi.org/10.1021/ct300418h

92. Majumder S, Deganutti G, Pipitò L, Chaudhuri D, Datta J, Giri K (2023). Computer-aided de novo design and optimization of novel potential inhibitors of HIV-1 Nef protein. *Comput Biol Chem*.104:107871. https://doi:10.1016/j.compbiolchem.2023.107871

93. Negi SS, Schein CH, Oezguen N, Power TD, Braun W (2007). InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics,* 23(24): 3397–3399. https://doi.org/10.1093/bioinformatics/btm474

94. Zerbe BS, Hall DR, Vajda S, Whitty A, Kozakov D (2012). Relationship between hot spot residues and ligand binding hot spots in protein-protein interfaces. *J Chem Inf Model*. ,52(8):2236-2244. https://doi:10.1021/ci300175u

95. Sukhwal A; Sowdhamini R (2013). Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Mol. Biosyst*. 9: 1652-1661. https://doi.org/10.1039/C3MB25484D

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2023.2184634

Taylor & Francis
Taylor & Francis Group

Check for updates

In silico investigation of binding propensity of hematoxylin derivative and damnacanthal for their potential inhibitory effect on HIV-1 Vpr from different subtypes

Joyeeta Datta, Satyabrata Majumder, Dwaipayan Chaudhuri and Kalyan Giri

Department of Life Sciences, Presidency University, Kolkata, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

HIV-1, the causative agent of AIDS leads to many deaths worldwide though few options are available as therapeutics. To deal with the continuous mutation in the virus genome, requirement of new drugs is always there. Subtype variation plays a crucial role in case of HIV-1 therapeutics development. In this study, we want to investigate some pre examined molecules that can be effective for HIV-1 VPR. Inhibition of several protein–protein interactions with the small molecules will lead to identify some molecules as therapeutics other than the conventional drugs. We retrieved the sequences of different subtypes from the database and representative sequences were identified. Representative structures were modelled and validated using MD simulations. Forty molecules, showing anti Vpr activity *in vitro* were identified from literature survey and those were docked with each subtype representative structures. Two molecules a stable Hematoxylin Derivative (SHD) and Damnacanthal (D3), these were shown to be bind more effectively for all the subtypes. The stability of the protein and those two small molecule complexes were identified again with MD simulation followed by the binding energy calculation. Thus, these molecules can be thought as any option other than the conventional drug targeting HIV-1 Vpr.

ARTICLE HISTORY
Received 8 December 2022
Accepted 20 February 2023

KEYWORDS
HIV-1 Vpr subtypes; MD simulation; damnacanthal; stable hematoxylin derivative; MM/PBSA binding energy

1. Introduction

Acquired Immuno Deficiency Syndrome (AIDS) remains one of the leading reasons of death globally. Along with the immune dysfunction, many other underlying reasons also affect the individuals suffering from AIDS. Two types of primate lentiviruses; Human Immunodeficiency Virus (HIV-1 and HIV-2) and Simian Immunodeficiency Virus (SIV) transmit through liquid contact from one host to another (Shaw & Hunter, 2012). HIV-1 remains the predominant causative agent of AIDS, however very few effective therapeutic interventions are available against the virus (Barouch, 2008; Dunne, 2007). There is always a requirement for novel, new therapeutics to combat the viral disease due to the high mutation rate and this process becomes more complicated for HIV-1 as they develop to form several groups, subtypes and Circulatory Recombinant Forms (CRF) (Chaplin et al., 2011; Gatell, 2011; Golmohammadi et al., 2017; Kantor et al., 2005; Lu et al., 2017; Rawson et al., 2015; Taylor et al., 2008; Wensing et al., 2019). Drugs available against HIV-1 can be mainly categorised in five different classes: Reverse transcriptase inhibitors, Integrase inhibitors, Fusion inhibitors, Protease inhibitors, Co receptor antagonist (Adamson & Freed, 2009; Arts & Hazuda, 2012). Exhaustive literature survey suggests that the infection efficiency, latency and viral pathogenesis vary with subtypes. Similarly, there are reports which confirm that the currently available drugs failed to provide cross subtype protection (Mazzuti et al., 2020; Neubert et al., 2016). Our motive of this study involves the investigation of subtype specific variation in the viral protein major binding pocket and the effect of the same on small molecule binding. This study may pave the way for development of new therapeutics utilizing a generalized approach so that the same molecule can be used for all the subtypes.

HIV-1 is a retrovirus with non-covalently linked two copies of positive sense single stranded RNA genome and encodes a total of 9 proteins which are essential for the entire viral life cycle (German Advisory Committee Blood (Arbeitskreis Blut), Subgroup 'Assessment of Pathogens Transmissible by Blood', 2016). Three structural proteins: Gag protein, Pol protein and Env protein are required for viral encapsidation and integration (Engelman & Cherepanov, 2012). Tat and Rev, two essential regulatory proteins are required for viral infection progression (Addo et al., 2001). There are four accessory proteins present in the HIV-1 genome namely Vif, Vpu, Vpr and Nef. Though these proteins don't have role in viral growth, several studies have reported about their role in virulence (Strebel, 2013; Zotova et al., 2019). Deficiency of these proteins leads to functional defects in the virus. As new strategies are required for development of new therapeutics, there are several researches going on by targeting the HIV-1 accessory proteins as drug target (Miller & Sarver, 1995). Therefore, we selected one of the accessory proteins

CONTACT Kalyan Giri kalyan.dbs@presiuniv.ac.in Department of Life Sciences, Presidency University, 86/1 College Street, Kolkata 700073, India
© 2023 Informa UK Limited, trading as Taylor & Francis Group

Certificate of Participation

This is to certify that

Joyeeta Datta

Presidency University, Kolkata, India

delivered *ORAL* presentation at the two day **International Conference (Online) on**

"Biomolecules To Biome" held on August 24-25, 2022

Organized by **Department of Life Sciences, Presidency University, Kolkata, India.**

BIOMOLECULES TO BIOME

N-terminal LIFE SCIENCES C-terminal

Dr. Ranjana Pal & Dr. Puja Ray
Joint Organizing Secretaries
Biomolecules To Biome - 2022

Prof. Prabir Kumar Mukhopadhyay
HOD (Life Sciences)
Presidency University, Kolkata
(Chairman)

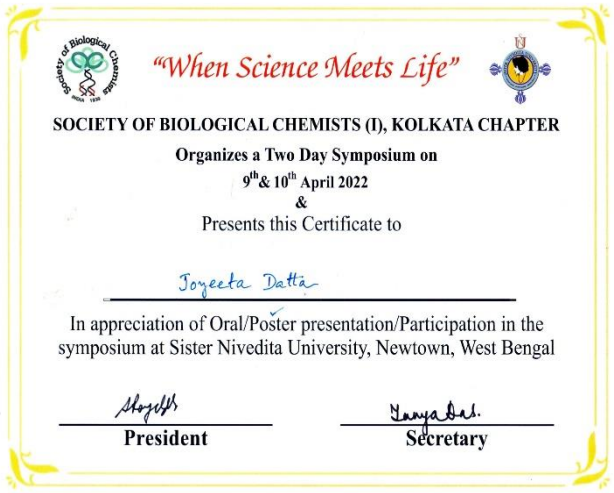
Prof. Saukar Bose
Dean of Faculty of Natural and
Mathematical Sciences
(Patron)

4th REGIONAL SCIENCE AND TECHNOLOGY CONGRESS
(SOUTHERN REGION), 2019
Jointly Organised By
DEPARTMENT OF SCIENCE & TECHNOLOGY AND
BIO-TECHNOLOGY, GOVT. OF WEST BENGAL
&
MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL

This is to certify that Prof./Dr./Mr./Ms. Joyeeta Datta of Presidency University is the presenting Author / Co-author / Corresponding Author in Oral / Poster / Exhibition session of the 4th Regional Science and Technology Congress (Southern Region) 2019, jointly organised by Department of Science and Technology and Bio-Technology, Government of West Bengal & Maulana Abul Kalam Azad University of Technology, West Bengal during December 23-24, 2019 at Haringhata, Nadia-741249.

Dr. Somdatta Chakravortty
Organising Secretary
4th Regional Science and Technology Congress (Southern Region), 2019
Maulana Abul Kalam Azad University of Technology, West Bengal

Dr. Suchandra Choudhury
Organising secretary
4th Regional Science and Technology Congress (Southern Region), 2019
Department of Science & Technology and Bio-Technology, Govt. of West Bengal



"When Science Meets Life"

SOCIETY OF BIOLOGICAL CHEMISTS (I), KOLKATA CHAPTER

Organizes a Two Day Symposium on
9th & 10th April 2022
&
Presents this Certificate to

Joyeeta Datta

In appreciation of Oral/Poster presentation/Participation in the symposium at Sister Nivedita University, Newtown, West Bengal

President

Secretary

DrillBit

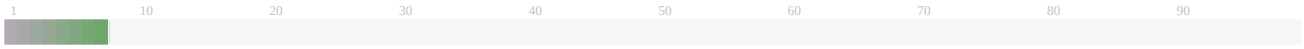
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

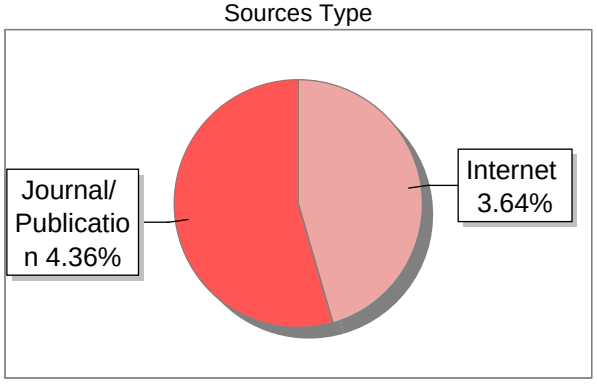
Author Name	Joyeeta Datta
Title	Studies on interactions of HIV-1 accessory proteins from different subtypes and their interacting partners
Paper/Submission ID	2287988
Submitted by	kalyan.dbs@presiuniv.ac.in
Submission Date	2024-09-05 17:38:09
Total Pages, Total Words	15, 8363
Document type	Synopsis

Result Information

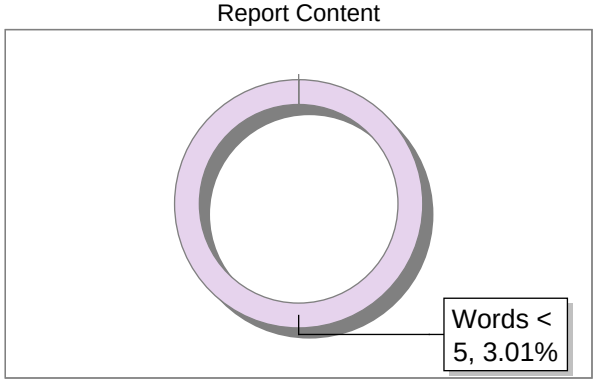
Similarity **8 %**



Sources Type



Report Content



Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 5 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File